ProUCL
USER'S GUIDE

Version 2.1

FEBRUARY, 2003

Table of Content

Installation Instructions

- Log in with administrator privileges. This is necessary to install and register controls. This is needed for the initial installation only. Subsequent updates will not require administrator privileges.

- Caution: This upgrade will install over previous version. If you wish to retain previous versions of ProUCL you will need to manually rename that directory and manually delete the existing ProUCL icon.

- It is strongly recommended that you exit all windows programs before running the setup program.

- Run the setup program contained on the CD.

- When prompted, accept the default directory or select another one of your choice.

Minimum Hardware Requirements

- Intel Pentium 200MHz

- 10 MB of hard drive space

- 48 MB of memory (RAM)

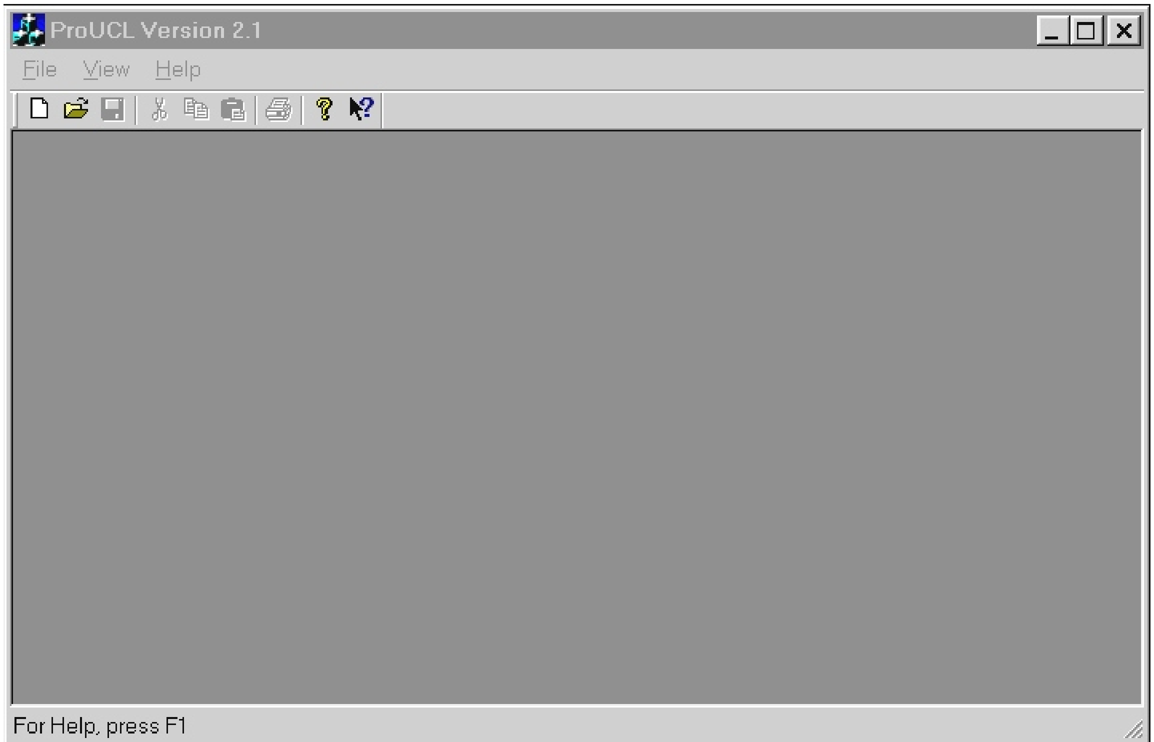
- CD-ROM drive

- Windows 98

Program ProUCL Menu Structure

The menu structure of ProUCL is similar to a typical Widows program. The screen below appears when the program is executed.

The following menu options appear on the screen



1. File

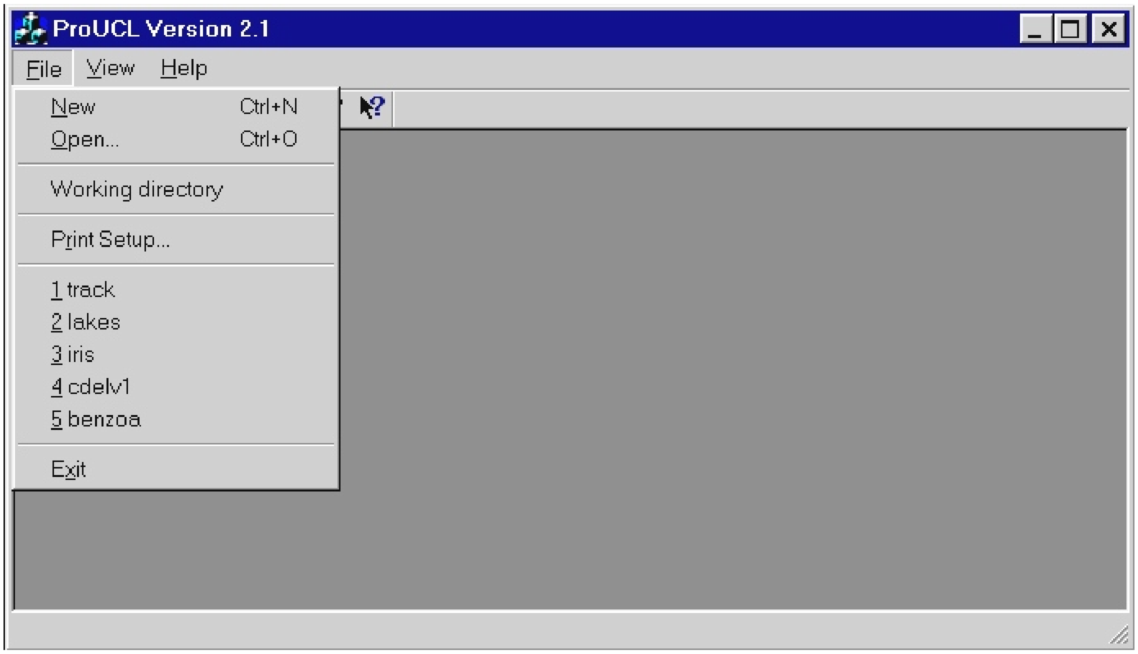
2. View

3. Help

The options available with these menu items are described next.

1. File

Click on the File menu item to reveal these drop-down options.



The following File drop-down menu options are available:

- New option: creates new spreadsheet.

- Open option: browses the disk for a file. The browse program will start in the working directory if a directory has been set.

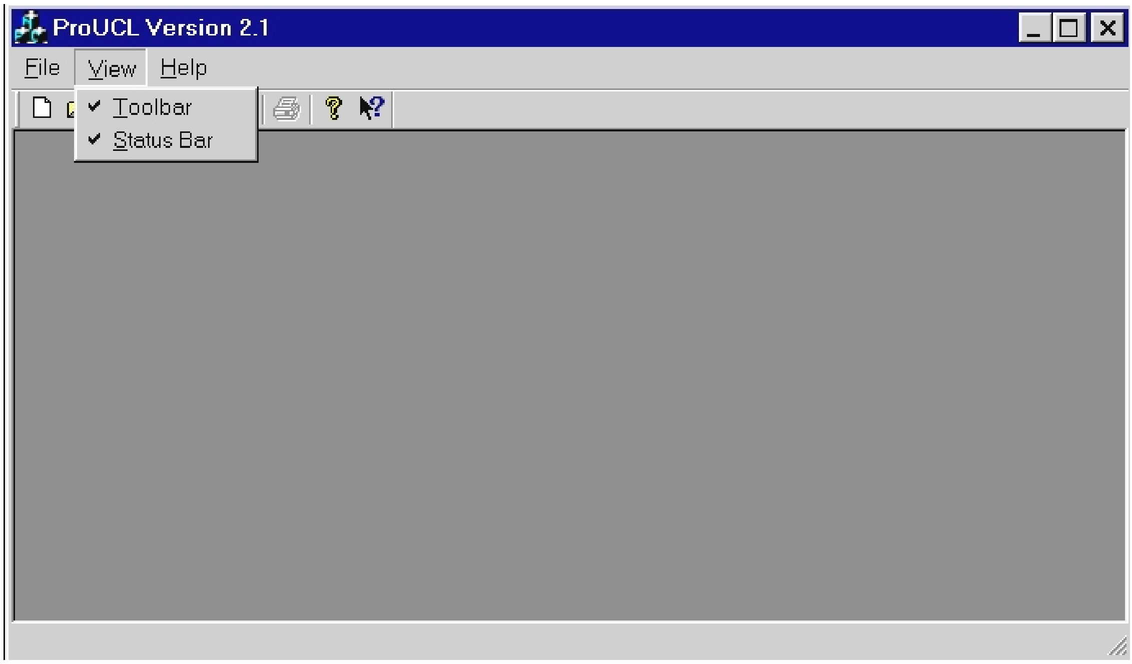
- Working directory option: select and set a working directory. Note: A file within a directory must be selected before setting the directory. All subsequent files are read from and saved in the chosen working directory.

- Print Setup option: sets printer options.

- Click on a previously used file to re-open that file.
- Exit option: exits ProUCL.

2. View

Click on the View menu item to reveal these drop-down options.



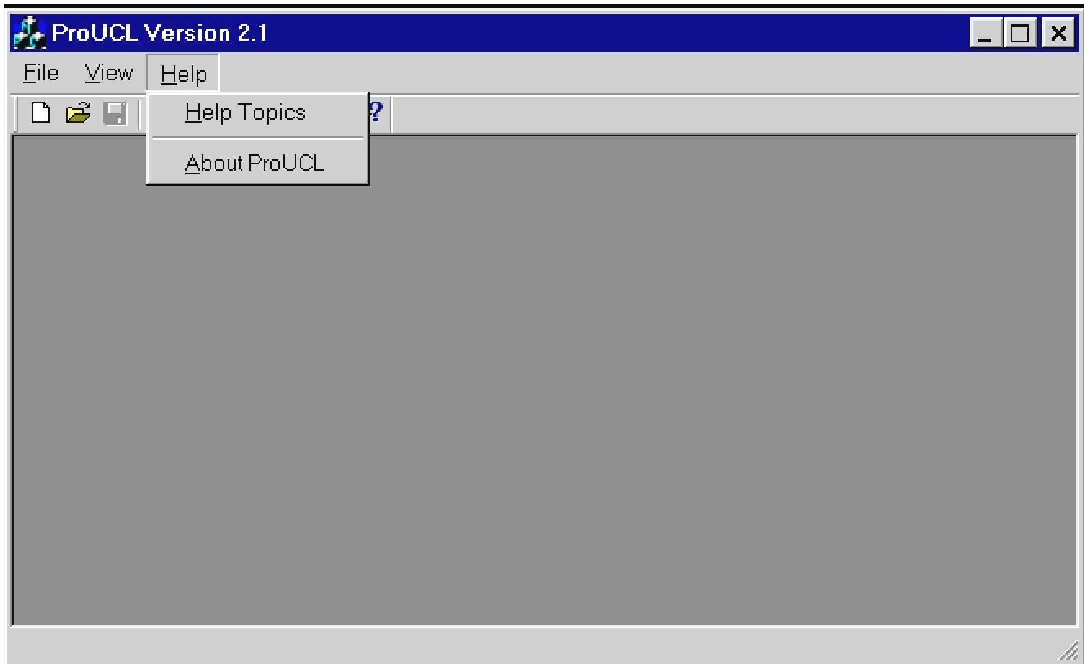
The following View drop-down menu options are available:

- Toolbar: the Toolbar is that row of symbols immediately below the menu items. Clicking on this option toggles the display. This is useful if the user wants to view more data on the screen.

- Status Bar: the Status Bar is the wide bar at the bottom of the screen which displays helpful information. Clicking on this option toggles the display. This is useful if the user wants to view more data on the screen.

3. Help

Click on the Help menu item to reveal these drop-down options.

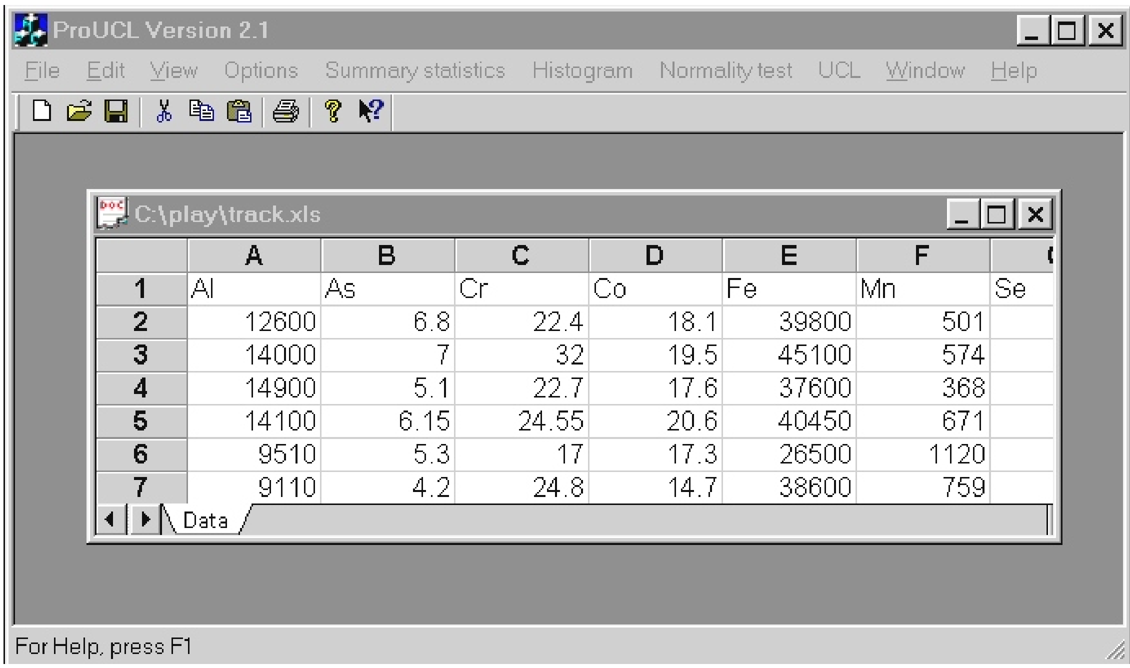


The following Help drop-down menu options are available:

- Help Topics option: at present no online help is available. This may be available in the next version of ProUCL.
- About ProUCL: displays the program version number.

Main Menu Structure of ProUCL

The following menu structure of ProUCL appears after opening or creating a data file.



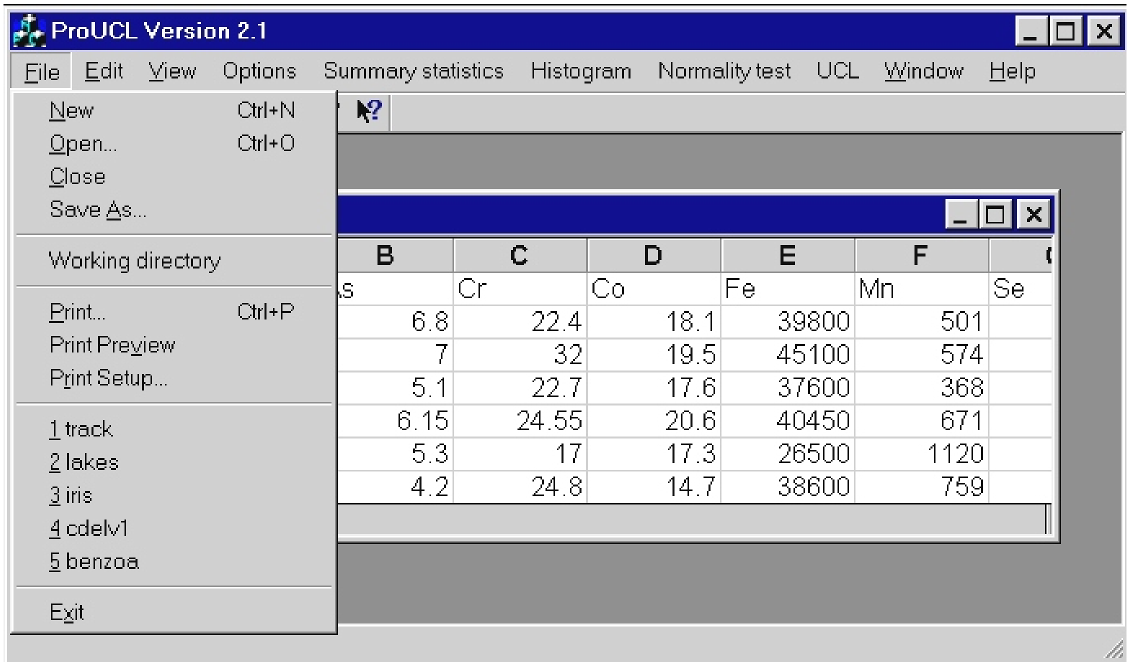
The following menu items are available.

1. File
2. Edit
3. View
4. Options
5. Summary statistics
6. Normality test
7. UCL
8. Window
9. Help

The options available with these menu items are described next.

1. **File**

Click on the File menu item to reveal these drop-down options.
The following File drop-down menu options are available:



- New option: opens a blank spreadsheet screen.

- Open option: browses the disk and selects a file which is then opened in spreadsheet format. The browse program will start in the working directory if a working directory has been set.

  Recognized input format options:

Text	*.txt (tab delimited)
Excel	*.xls
Lotus	*.wk?
Lotus	*.123
Default -	*.* will be read in Excel format.

- Close option: closes the active window.
- Save As option: allows the user to save the active window.

Follows the Windows standard and writes to a file in Excel 95 format. All modified/edited data files, and output screens generated by the software, can be saved in Excel 95 format.

- Working directory option: selects and sets a working directory for all I/O actions. All subsequent files are read from and saved in the working directory. <u>You must select a file before you set the working directory.</u>

- Print option: sends the active window to the printer.

- Print Preview option: displays a preview of the output on the screen

- Print Setup options: follow Windows standard. The user can chose the landscape format under this option.

<u>Input File Format</u>

- The program can read Tab delimited Text (ASCII), Excel, and Lotus files.

- Columns in a Text (ASCII) file should be separated by one tab. <u>Spaces between columns are not allowed in this format.</u>

- The input data file should have column labels in the first row and data without text (e.g., non-numeric and blank values) for those variables in the remaining rows.

- The data file can have multiple variables (columns) with unequal number of observations.

- Non-numeric text may only appear in the header row (first row) of each column. All other non-numeric data (blank, other characters, and strings) appearing elsewhere in the data file are treated as zero entries. The user should make sure that his data set does not contain such non-numeric values.

- Alternatively, a large value = 1E31 ($=1\times10^{31}$) can be used for missing (blank, or non-numeric values) observations (just as in Scout (1999) software). All values with this large value are ignored from all of the computations.

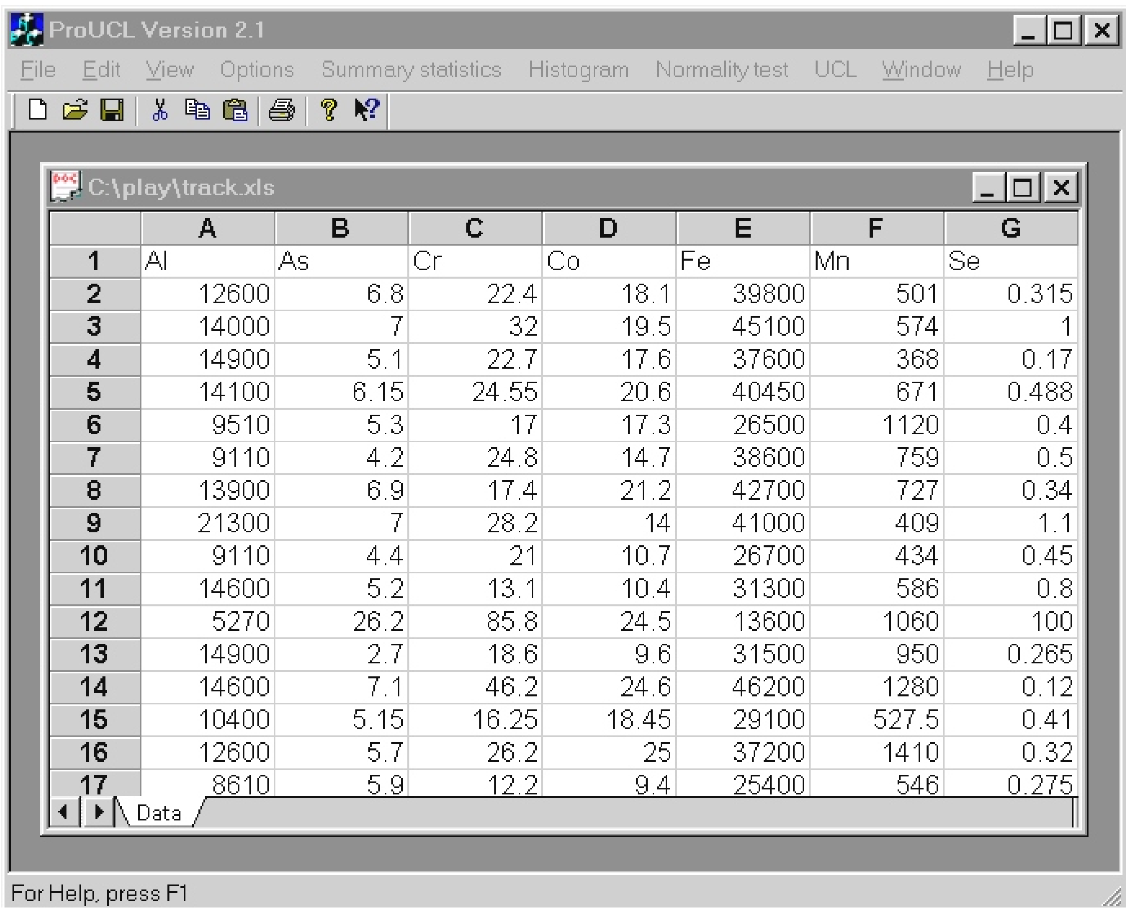
- Data in each column must end with a <u>non-zero value.</u> The last non-zero entry in each column is considered as the end of that column's data. If your data column ends with a zero value, that last zero value will be ignored. This may require you move observations around if your column data ends with zero values.

- Note that all other zero data (in the beginning or middle of a data column) are treated as valid zero values.

- At present, the program does not handle the left-censored data sets with non-detects.

Result of Opening an Input Data File

- The data screen follows the standard Windows design. It can be



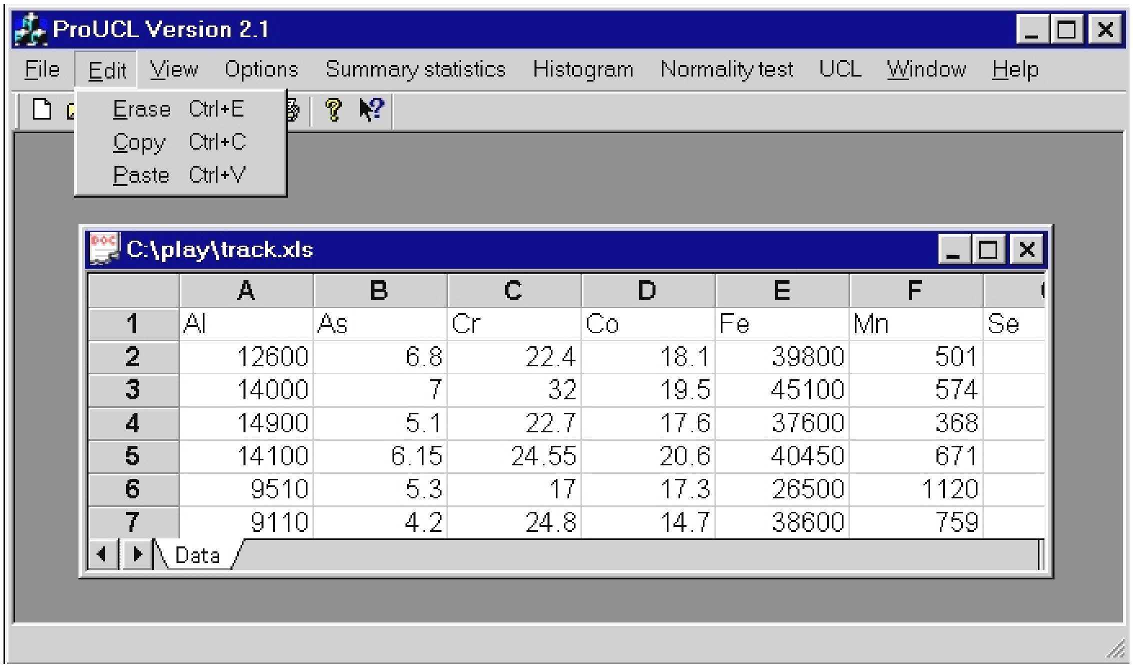
	A	B	C	D	E	F	G
1	Al	As	Cr	Co	Fe	Mn	Se
2	12600	6.8	22.4	18.1	39800	501	0.315
3	14000	7	32	19.5	45100	574	1
4	14900	5.1	22.7	17.6	37600	368	0.17
5	14100	6.15	24.55	20.6	40450	671	0.488
6	9510	5.3	17	17.3	26500	1120	0.4
7	9110	4.2	24.8	14.7	38600	759	0.5
8	13900	6.9	17.4	21.2	42700	727	0.34
9	21300	7	28.2	14	41000	409	1.1
10	9110	4.4	21	10.7	26700	434	0.45
11	14600	5.2	13.1	10.4	31300	586	0.8
12	5270	26.2	85.8	24.5	13600	1060	100
13	14900	2.7	18.6	9.6	31500	950	0.265
14	14600	7.1	46.2	24.6	46200	1280	0.12
15	10400	5.15	16.25	18.45	29100	527.5	0.41
16	12600	5.7	26.2	25	37200	1410	0.32
17	8610	5.9	12.2	9.4	25400	546	0.275

resized, or portions of data can be viewed using scroll bars.

- Note that scroll bars appear when the window is activated and the title bar is highlighted.

2. Edit

Click on the Edit menu item to reveal drop-down options.
The following Edit drop-down menu options are available:



- Erase option: is used to remove the highlighted portion of the data. Note that the erased data is not written to any buffer and cannot be recovered. Therefore, when erased, it is gone.

- Copy option: is similar to a standard Windows Edit option such as in Excel. It performs typical edit functions of copying highlighted data to a buffer.

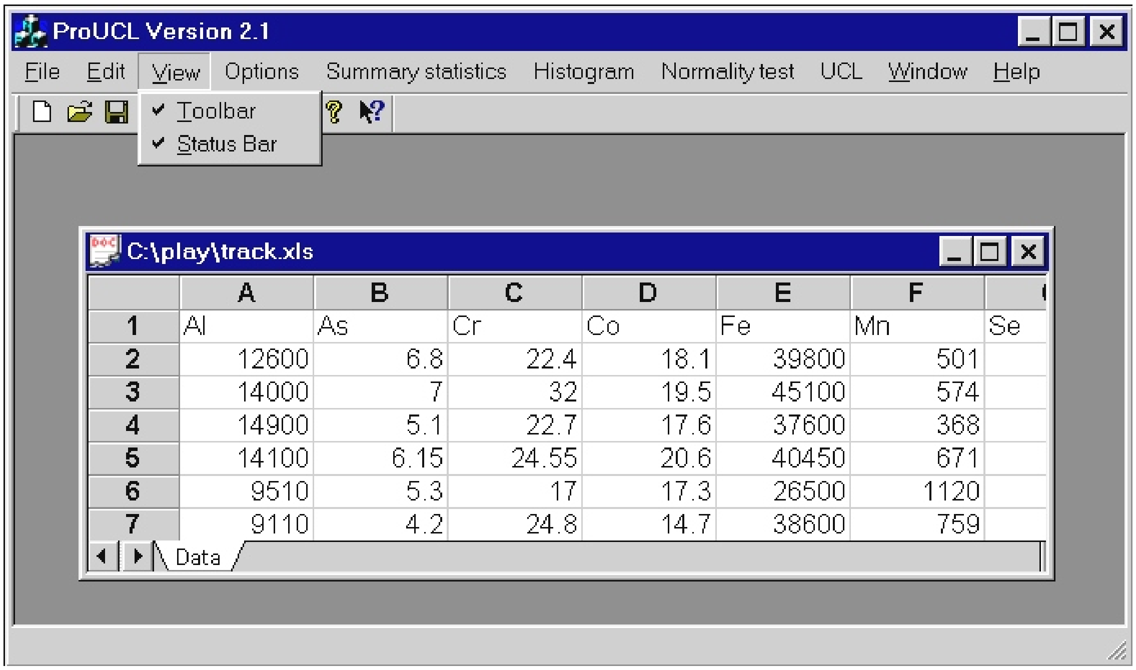
- Paste option: is similar to a standard Windows Edit option such as in Excel. It performs typical edit functions of pasting data from a buffer to the current spreadsheet cell.

3.View

Click on the View menu item to reveal these drop-down options.

The following View drop-down menu options are available:

- Toolbar: the Toolbar is that row of symbols immediately below

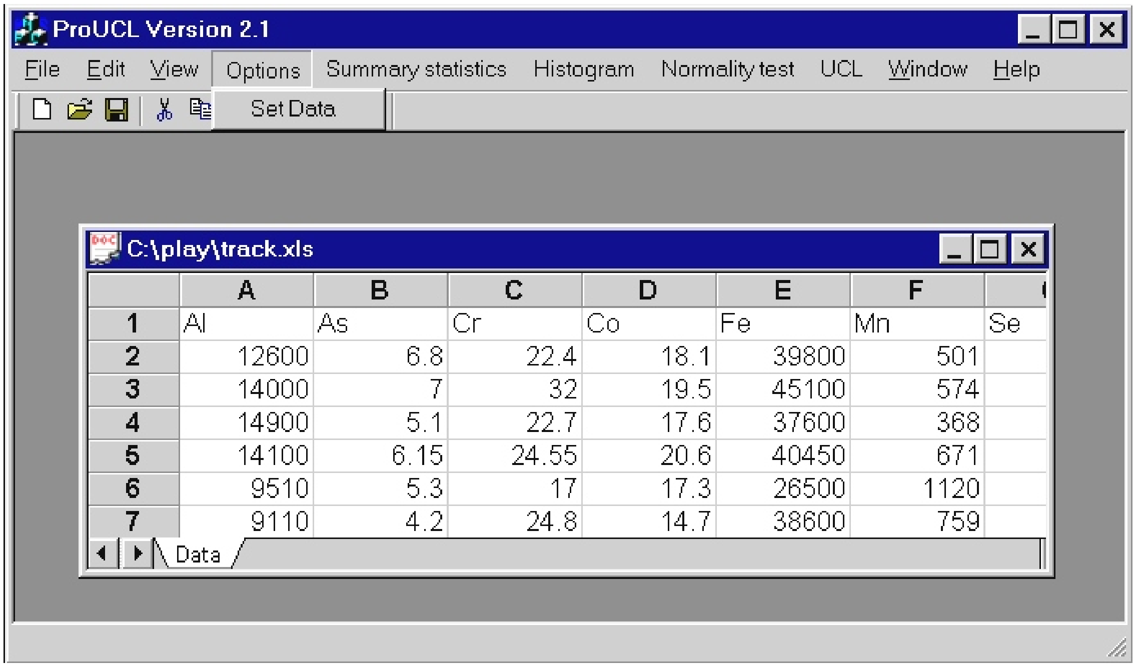


the menu items. Clicking on this option toggles the display. This is useful if the user wants to view more data on the screen.

- Status Bar: the Status Bar is the wide bar at the bottom of the screen which displays helpful information. Clicking on this option toggles the display. This is useful if the user wants to view more data on the screen.

4. Options

Click on the Options menu item to reveal these drop-down options.



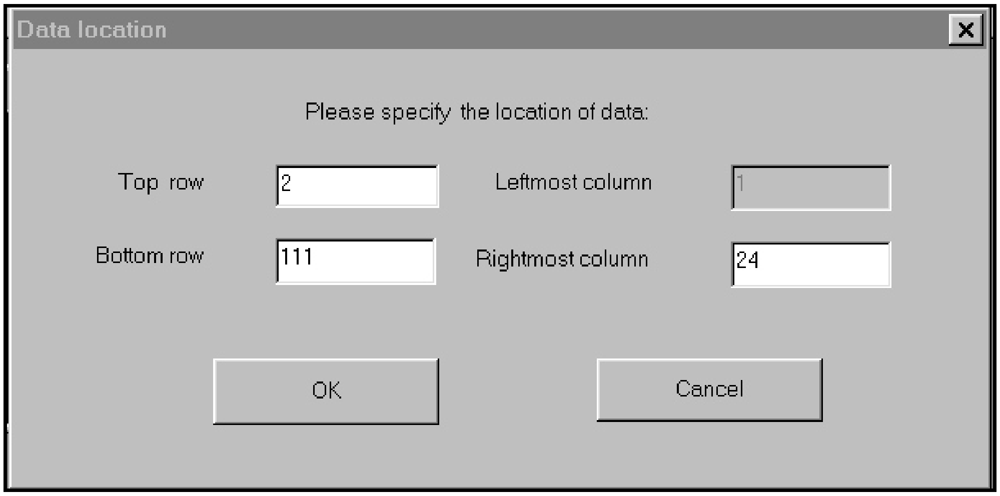
Currently, Set Data is the only drop-down menu option available:

- Set Data option: resets the active portion of the data window. The program examines the active spreadsheet and selects default values representing the first row of data (row 2), the last row which contains data (dependent on actual data), the leftmost column (typically column 1) where data and text occur, and the rightmost column (dependent on actual data) where data and text occur. Extreme caution should be taken when varying from the default values.

- **Note: This menu item is optional. The user can pre-process the data by using the Excel program.**

The Data Location Screen

The following Data location screen appears.



- **It is recommended to use the default settings for the data screen. This means that all of the data will be processed.**

- The first row in the spreadsheet contains the alphanumeric text (column headings), not data.

- The default top row of data is row 2. This value can be changed to process a subset of the data in the spreadsheet.

- The default bottom row is the last row in the spreadsheet which contains nonzero data. This value can be changed to process a subset of the data in the spreadsheet.

- The selected data must correspond to the same columns as the text in the first row. The Leftmost column value (column number) cannot be changed by the user.

- Caution: it is possible to confuse the program by highlighting a portion of the spreadsheet before invoking this option, unpredicted results will occur.

- The Rightmost column number can be changed by the user. Note that you must have a column of data for any variable requested.

- Caution: Blank cells in the top data row may confuse the automatic sizing algorithm. The user can manually override this confusion by re-setting the rightmost column value in this option.

5. Summary statistics

- This option computes general summary statistics for **all variables** in the data file.

- Two Choices are available:

  Raw data (the default option)
  Log-transformed data (Natural logarithm)

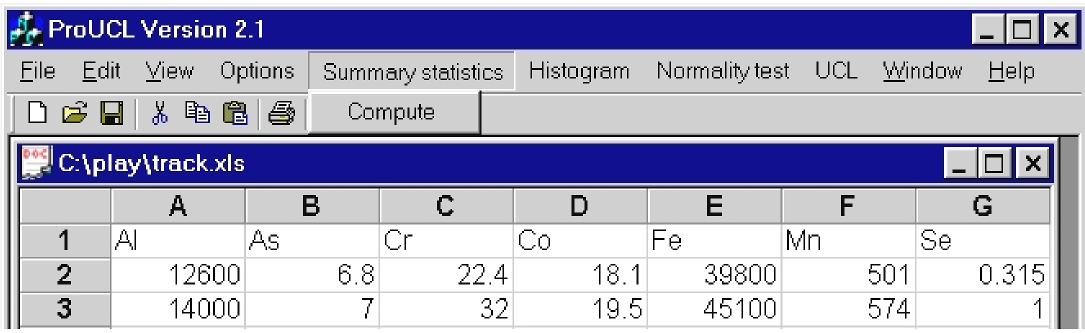
- In ProUCL, Log-transform means natural logarithm (ln).

- When computing summary statistics for raw data, an informative message is displayed for each variable which may contain non-numeric or non-positive values.

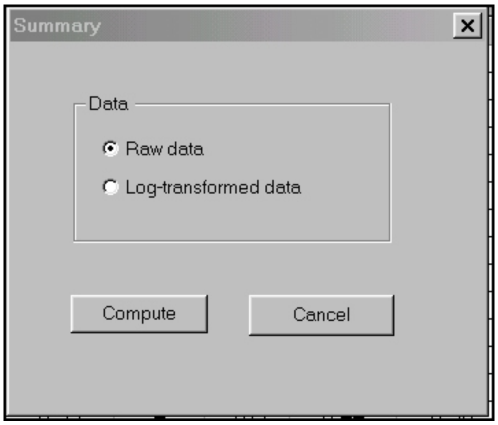
- The Summary statistics option computes log-transformed data only if all of the data values for the selected variable are positive real numbers. A message will be displayed if non-numeric characters, zero, or negative values are found in the column corresponding to the selected variable.

Summary Statistics

Click on the Summary statistics menu item to reveal this drop-down option.



When the user clicks on the Compute option button, the window given on the right appears.

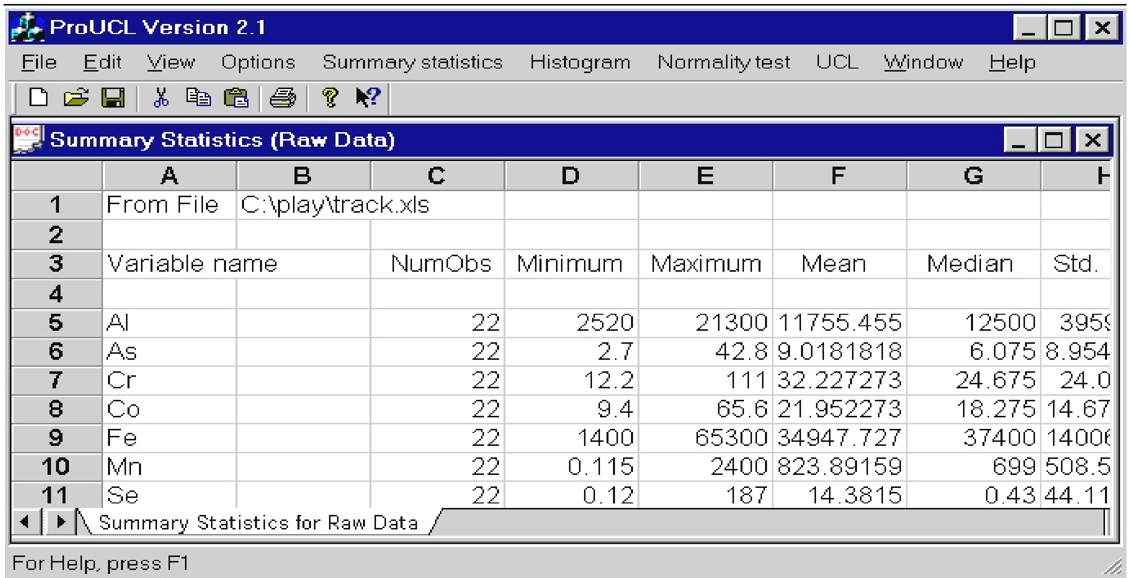


- Select your data choice, and click on the Compute button to continue or on the Cancel button to cancel the summary operations.

- The results screen follows the standard Windows design. It can be edited, widened, printed, resized, or scrolled.

- The resulting summary statistics screen can be saved as an Excel file.

Results Obtained Using the Summary Statistics Option



	A	B	C	D	E	F	G	H
1	From File	C:\play\track.xls						
2								
3	Variable name		NumObs	Minimum	Maximum	Mean	Median	Std.
4								
5	Al		22	2520	21300	11755.455	12500	395
6	As		22	2.7	42.8	9.0181818	6.075	8.954
7	Cr		22	12.2	111	32.227273	24.675	24.0
8	Co		22	9.4	65.6	21.952273	18.275	14.67
9	Fe		22	1400	65300	34947.727	37400	1400
10	Mn		22	0.115	2400	823.89159	699	508.5
11	Se		22	0.12	187	14.3815	0.43	44.11

On the results screen, the following summary statistics are displayed for each variable in the data file. These are described in Appendix A.

- NumObs - Number of Observations.
- Minimum - Minimum value.
- Maximum - Maximum value.
- Mean - Average value.
- Median - Median value.
- Std. Dev.- Standard Deviation.
- CV - Coefficient of Variation.
- Skewness - Skewness statistic.
- Variance - Variance statistic.

Printing Summary Statistics

- The summary statistics results and all other results can be printed by clicking the Print option under the menu item File. It is recommended that these statistics be printed in landscape format which is available under the Print Setup option.

6. Histogram

- This option produces a histogram for a selected variable in the data file.

- For data sets with more than one variable, the user should select a variable first. **The histogram is computed and displayed for the selected variable, one variable at a time.**

  - By default, the program selects the first variable.

- The user specifies if the data should be transformed.

  - The default choice is to display a raw data histogram.

- Two Choices are available:

  Raw data (the default option)
  Log-transformed data (Natural logarithm)

- In ProUCL, Log-transform means natural logarithm (ln).

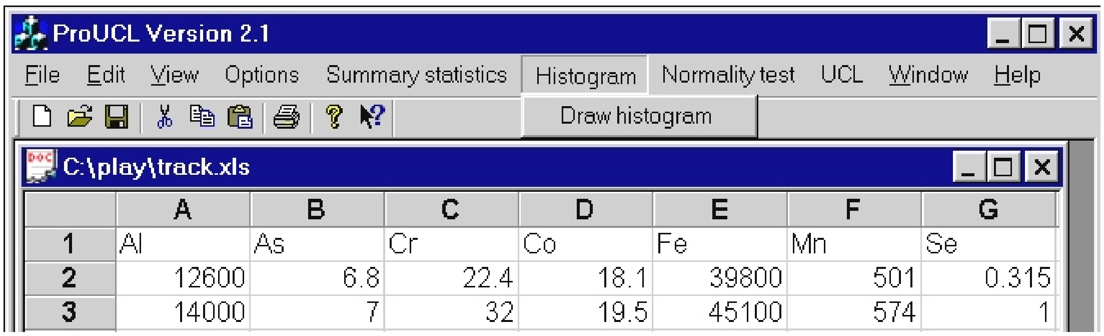
- The user can select the number of bins for the histogram.

  The default is 15 bins.

- Note that in order to display and capture the best histogram window, the user may want to **maximize the window before printing.**

Histogram Screen

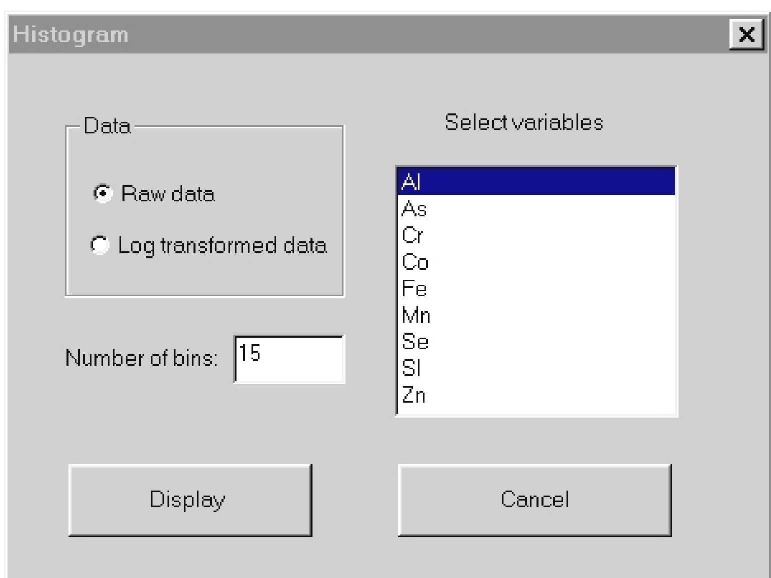
- Click on the Histogram menu item and select the Draw histogram option.



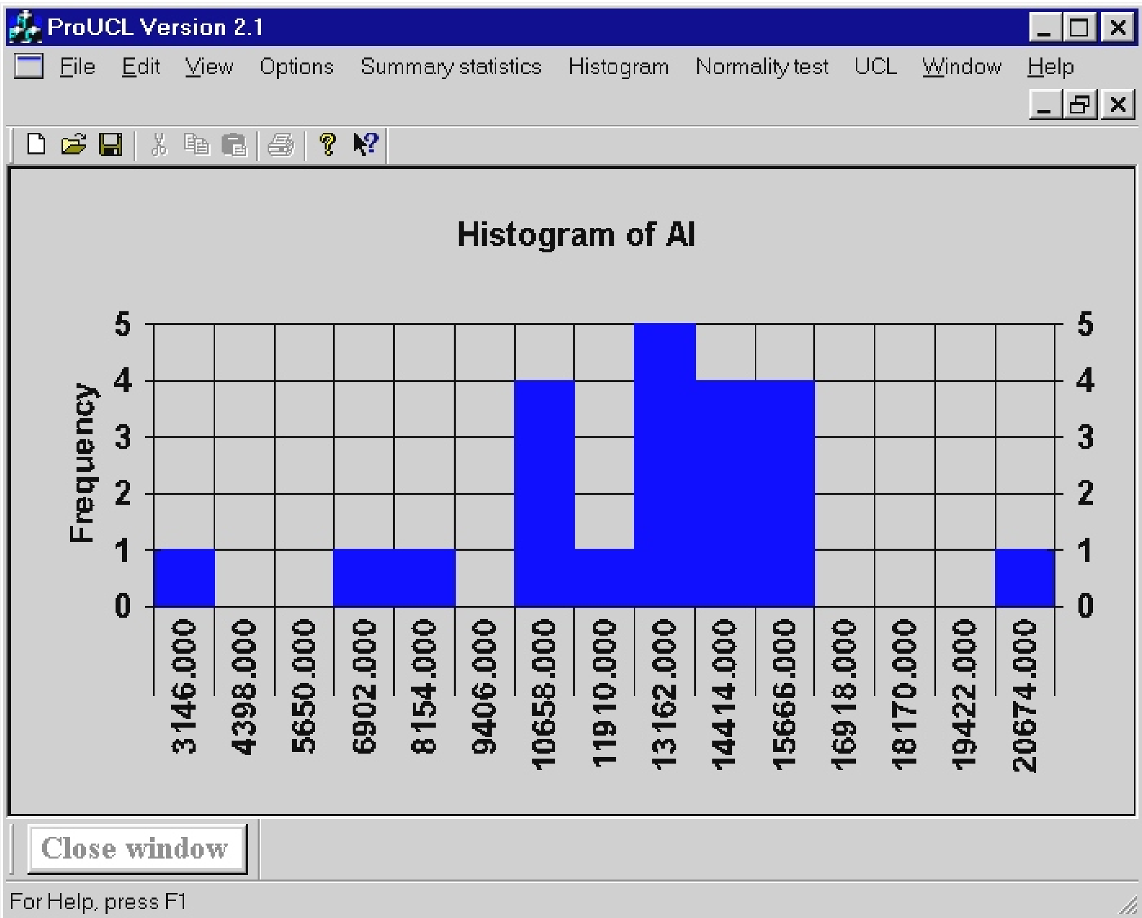
- Select Raw data or Log transformed data.

- You can change the number of bins to display.

- Select the variable you wish to view the histogram and then hit the display key to view the histogram.



Results of Histogram Option



- The Histogram window shown above has been resized for display and reflects the default values shown on the previous page.

- You may close the window using normal windows operations or click on the Close window button at the bottom left corner of the screen.

- The histogram can be printed or copied by clicking on the right button on mouse.

7. Normality test

- This option tests the normality or lognormality of the variable selected by the user.

- For data sets with more than one variable, the user should select a variable first. **The normality is tested and displayed for the selected variable, one variable at a time.**

  - By default, the program selects the first variable.

- The user specifies the transformation (normal or lognormal).

  - The default choice is to test for normality.

- The user specifies level of significance. Three choices are available for the level of significance: 0.01, 0.05, or 0.1

  - The default choice for level of significance is 0.05

- The program ProUCL plots a normal quantile-quantile (Q-Q) plot for the selected variable (or the log-transformed variable).

- The linear pattern of the Q-Q plot suggests approximate normality (or lognormality).

- The Program computes the intercept, slope, and correlation coefficient for the linear pattern displayed by the Q-Q plot. A high value (e.g., >0.95) of the correlation coefficient is an indication of approximate normality. Note that these statistics are among those displayed on the Q-Q plot.

- Typically, on this graph, observations well separated from the bulk (central part) of data are potential outliers needing further investigation.

- In addition to the graphical Q-Q plot, two more powerful procedures are also available to test the normality or lognormality of the data. These are:

  – Lilliefors Test: a test typically used for samples of larger size (> 50). When the sample size is greater than 50, the program defaults to the Lilliefors test. However, note that the Lilliefors test is available for samples of all sizes.

  – Shapiro and Wilk W-Test: a test used for samples of smaller size (<=50). At present, W-Test is available only for samples of size 50 or less.

- ProUCL computes the relevant test statistic and the associated critical value, and prints them on the associated Q-Q plot.

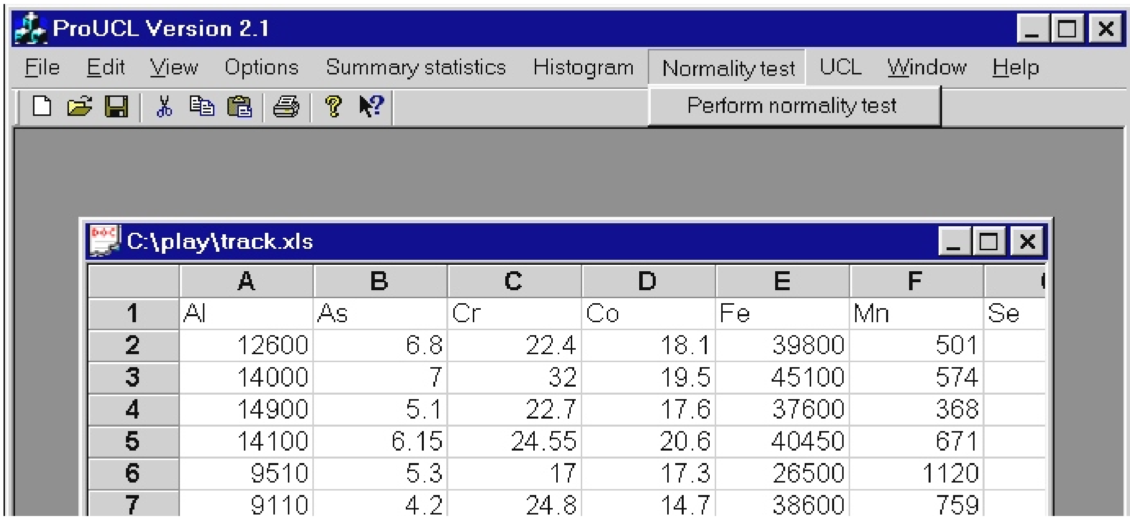
- On this Q-Q plot, the program informs the user if the data are normal (or lognormal).

- The Q-Q plot can be printed or copied by clicking the right button on the mouse.

- Note that in order to capture the entire graph window, the user may want to **maximize the window before printing.**

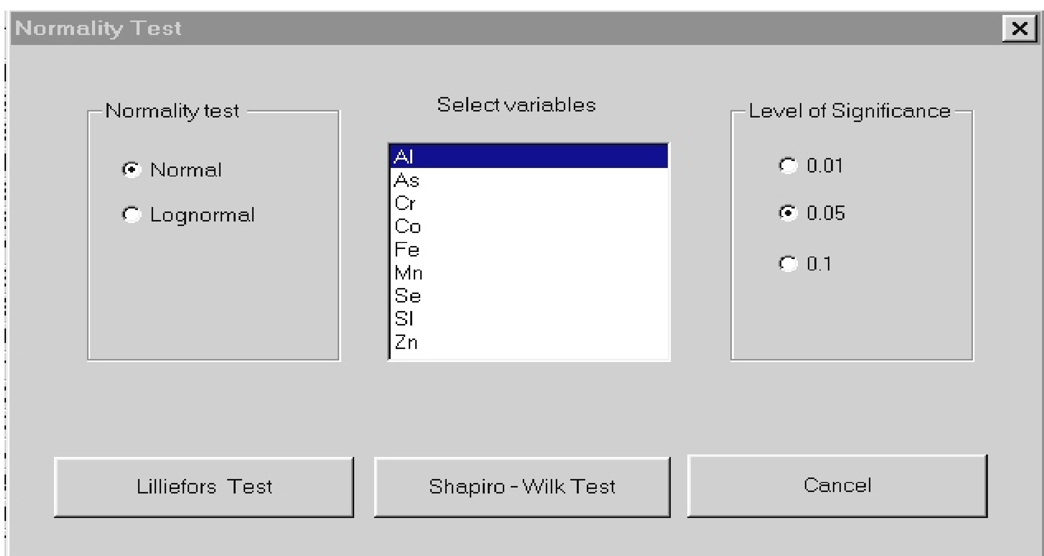
Normality test Screen

- Click on the Normality test menu item and select the Perform normality test option.



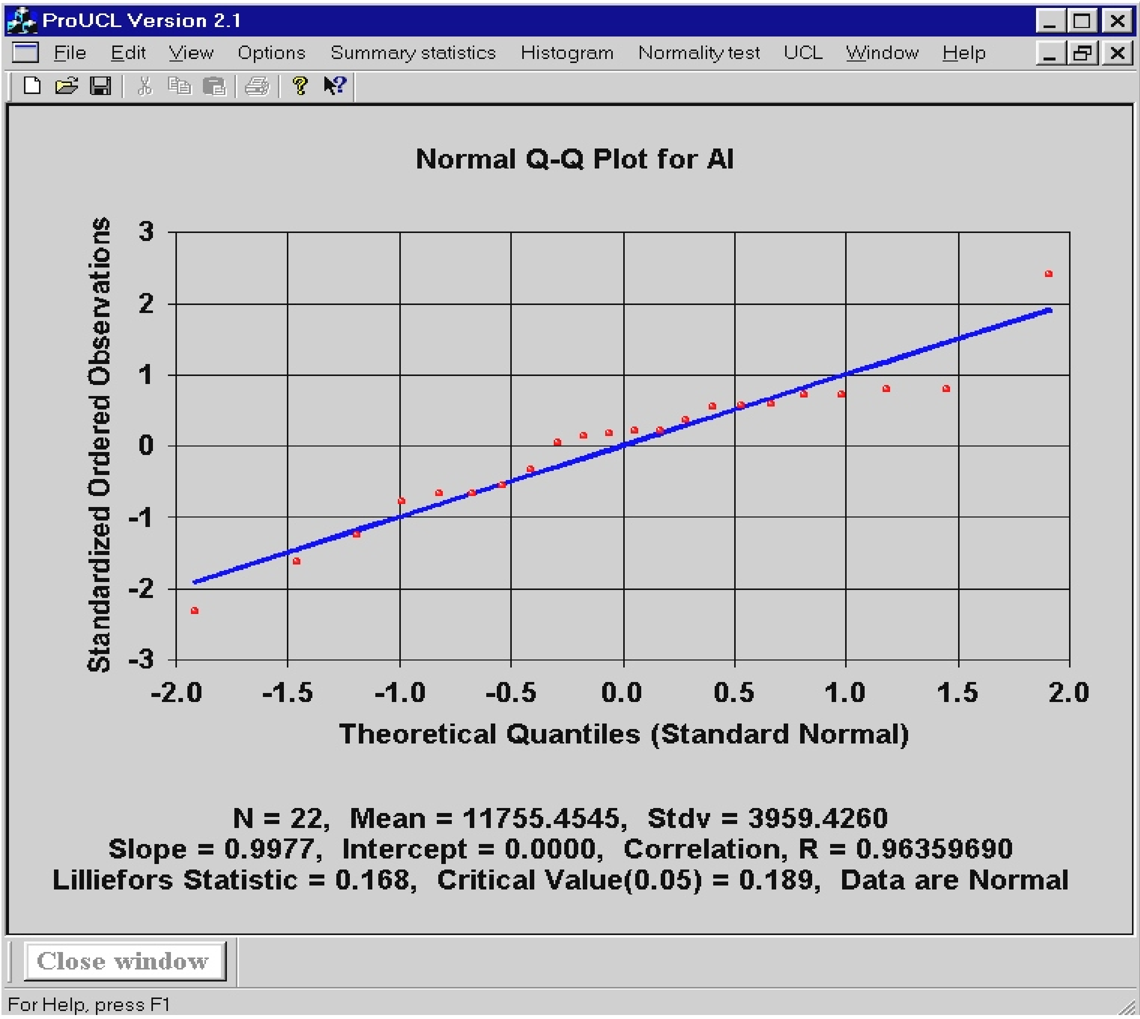
- Select either the Normal option or the Lognormal option.

- Select the variable, select a Level of Significance, and then click on the test (Lilliefors or Shapiro-Wilk) you wish to perform.



R **e s**

ults of Normality test Option



- The Q-Q plot window shown above has been resized for display.

- Two different Q-Q plot windows are produced for each Normality test request: using the original data (shown above) and the standardized data.

8. UCL

- This option computes the UCLs for the selected variable.

- This option allows the user to choose one or more methods from the several (10) available methods to compute a UCL of the population mean.

- By default, the program computes UCLs using all available methods.

- The user specifies the confidence level: a number in the interval (0.5, 1). The default choice is 0.95.

- The program computes several non-parametric UCLs using the Central Limit Theorem, Chebyshev inequality, Jackknife, and Bootstrap procedures.

- For the bootstrap method, the user can specify the number of bootstrap runs. The default choice for the bootstrap runs is 2000.

- The user is responsible for making an appropriate choice about data distributions - normal or lognormal. The user determines the data distribution using the normality test option. The program informs the user if the data are normal or lognormal. The program computes the relevant statistics using this choice.

- For data sets which are neither normal nor lognormal, ProUCL computes UCLs using non-parametric procedures.

- For lognormal data sets, ProUCL can compute only a 90% or a 95% H-statistic based H-UCL of the mean. For all other methods, it can compute a UCL for any confidence coefficient in the interval (0.5,1.0).

- For lognormal distributions, when the user wants to compute a 95% UCL, ProUCL also provides a recommended UCL computation procedure. This is particularly helpful when the skewness is high, that is, the standard deviation of the log-transformed data starts exceeding 1.

- For lognormal data sets, the program also computes the Maximum Likelihood Estimates (MLEs) of the population percentiles, and the minimum variance unbiased estimates (MVUEs) of population mean, median, standard deviation, and the standard error (SE) of the mean.

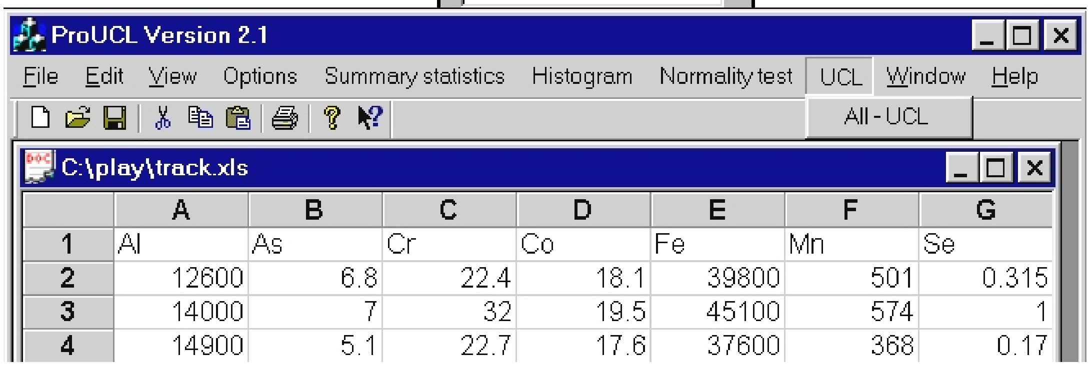
- ProUCL can compute the H-UCL for samples of size up to 1000 using the critical values as given by Land (1975).

- The detailed theory and formulae to compute these statistics are given by Land (1971, 1975), Gilbert (1987), Singh, Singh, and Engelhardt (1997, 1999), and Singh et al. (2000).

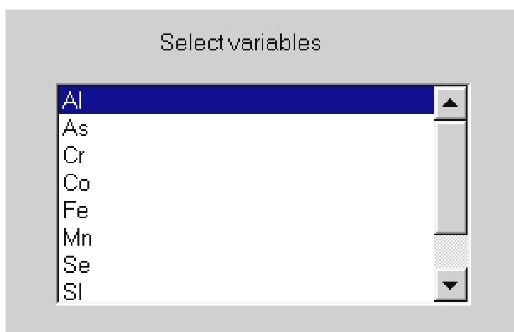
- For the sake of completeness of this User's Guide, all formulae and methods used in the development of the program ProUCL are summarized in Appendix A.

UCL Computation Screen

Click on the UCL menu item and then click on the All-UCL option.



- Note that the UCLs are computed for one variable at a time. The user selects a variable from the variable list.

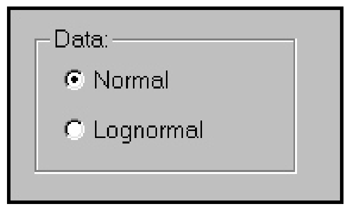


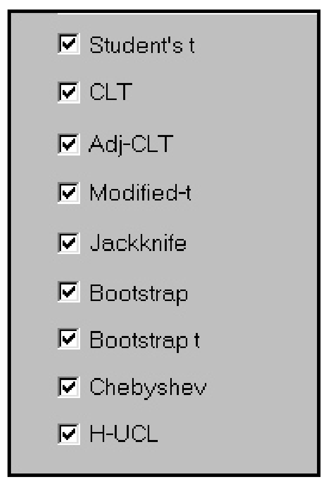
- The user may change the Confidence Coefficient (Default is 0.95). The range allowed is between 0.5 and 1.0.

- The user may adjust the number of bootstrap runs (Default is 2,000)

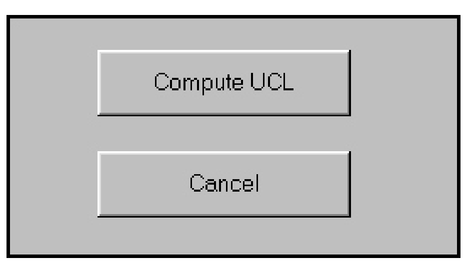
- The user selects the Normal or Lognormal Data option

- The user may de-select any unwanted UCL computations procedures.

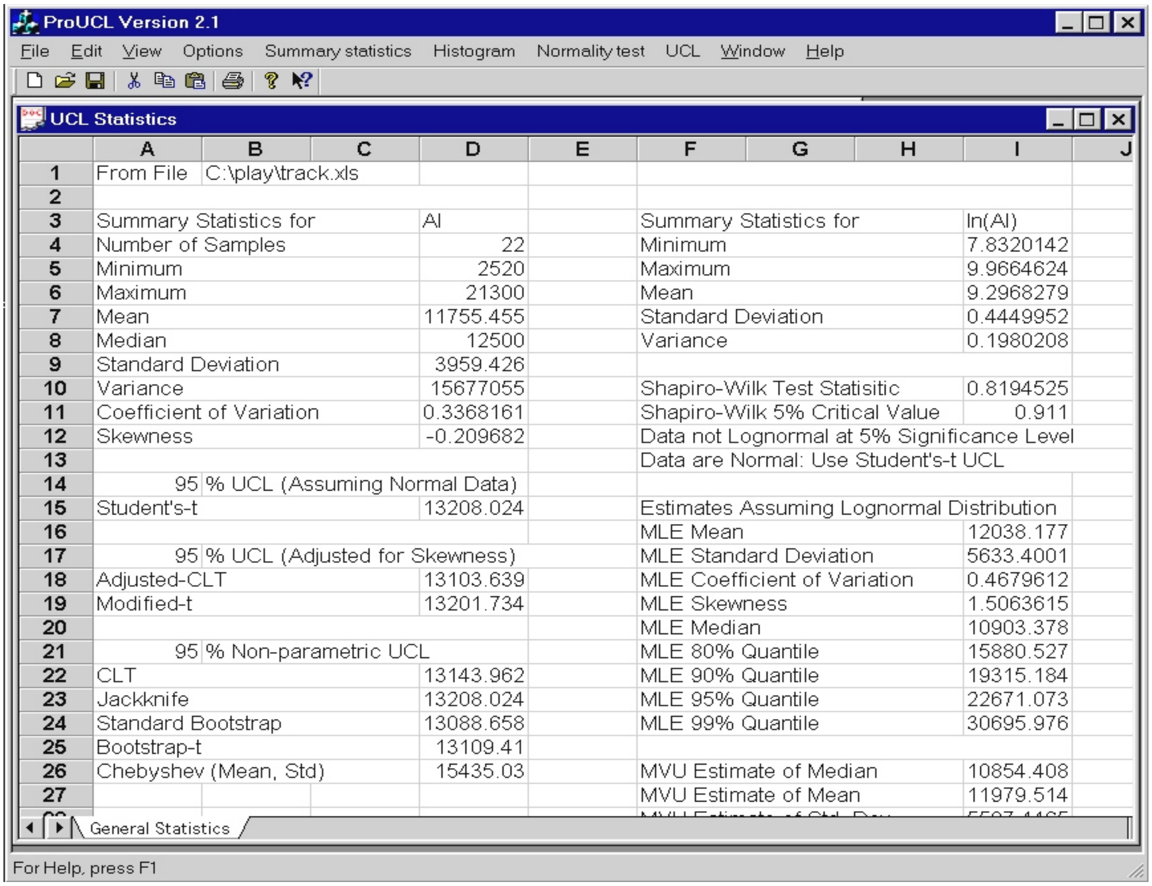




- Finally, the user clicks on the Compute UCL button.



Results Screen of UCL Computations



	A–C	D	F–H	I
1	From File C:\play\track.xls			
2				
3	Summary Statistics for	Al	Summary Statistics for	ln(Al)
4	Number of Samples	22	Minimum	7.8320142
5	Minimum	2520	Maximum	9.9664624
6	Maximum	21300	Mean	9.2968279
7	Mean	11755.455	Standard Deviation	0.4449952
8	Median	12500	Variance	0.1980208
9	Standard Deviation	3959.426		
10	Variance	15677055	Shapiro-Wilk Test Statisitic	0.8194525
11	Coefficient of Variation	0.3368161	Shapiro-Wilk 5% Critical Value	0.911
12	Skewness	-0.209682	Data not Lognormal at 5% Significance Level	
13			Data are Normal: Use Student's-t UCL	
14	95 % UCL (Assuming Normal Data)			
15	Student's-t	13208.024	Estimates Assuming Lognormal Distribution	
16			MLE Mean	12038.177
17	95 % UCL (Adjusted for Skewness)		MLE Standard Deviation	5633.4001
18	Adjusted-CLT	13103.639	MLE Coefficient of Variation	0.4679612
19	Modified-t	13201.734	MLE Skewness	1.5063615
20			MLE Median	10903.378
21	95 % Non-parametric UCL		MLE 80% Quantile	15880.527
22	CLT	13143.962	MLE 90% Quantile	19315.184
23	Jackknife	13208.024	MLE 95% Quantile	22671.073
24	Standard Bootstrap	13088.658	MLE 99% Quantile	30695.976
25	Bootstrap-t	13109.41		
26	Chebyshev (Mean, Std)	15435.03	MVU Estimate of Median	10854.408
27			MVU Estimate of Mean	11979.514

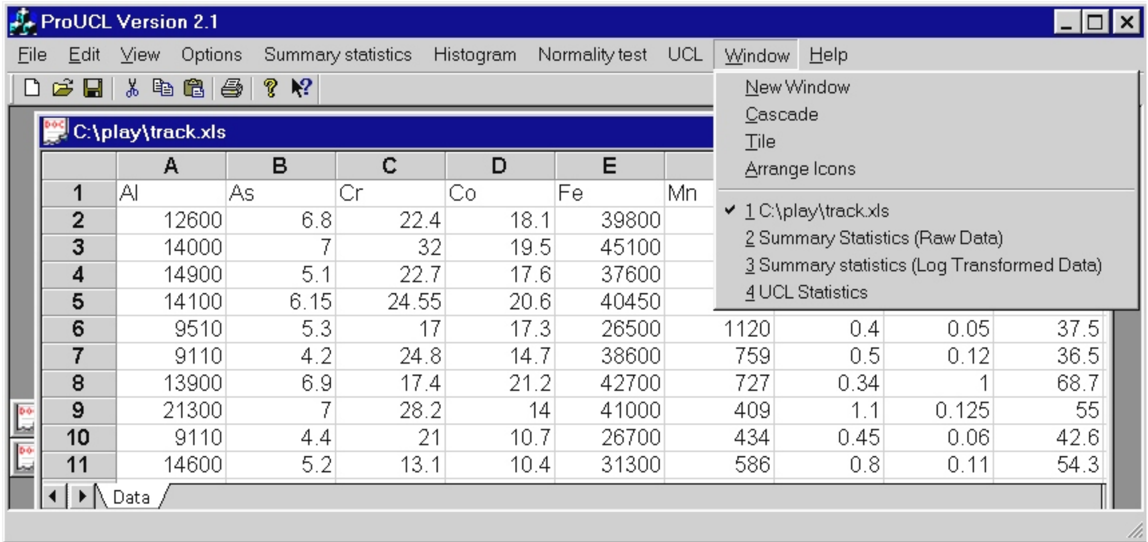
- On the output of ProUCL, Chebyshev (Mean, Std) stands for a Chebychev UCL of the mean computed using the sample arithmetic mean and standard deviation.

- 95% Chebyshev (MVUE) UCL stands for a 95% UCL of the mean obtained using the MVUEs of the mean and standard error of the mean assuming a lognormal distribution.

- 99% Chebyshev (MVUE) UCL stands for a 99% UCL of the mean obtained using the MVUEs of the mean and standard error of the mean assuming a lognormal distribution.

9. Window

Click on the Window menu to reveal these drop-down options,

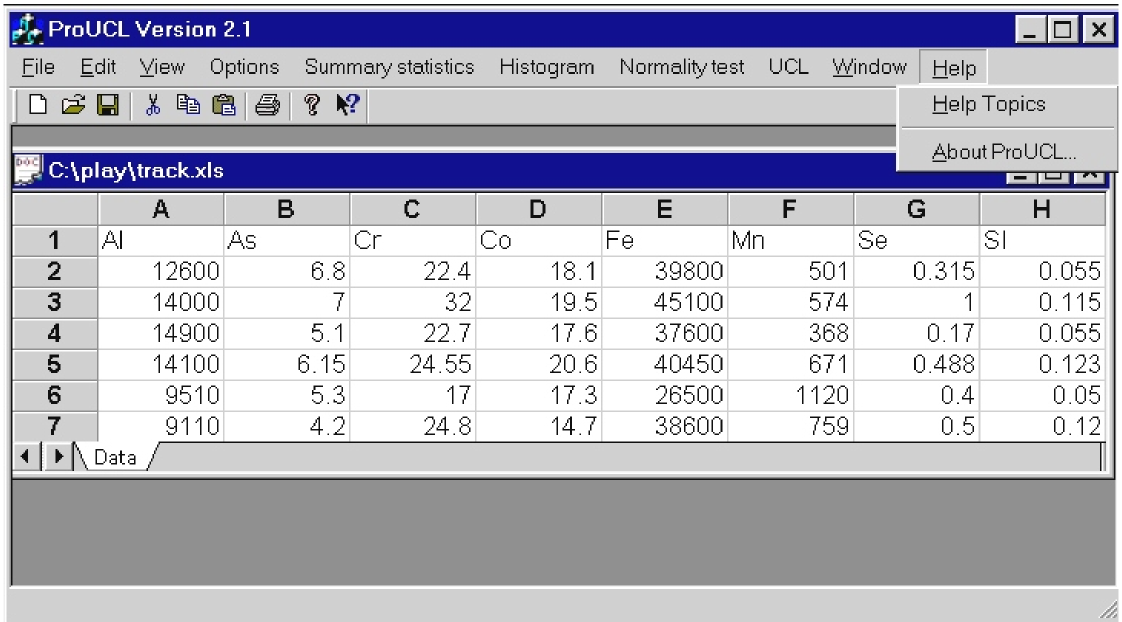


The following Window drop-down menu options are available:

- New Window option: opens a blank spreadsheet window
- Cascade option: arranges windows in a cascade format. This is a typical Windows program option.
- Tile option: resizes each window and then displays all open windows. This is a typical Windows program option.
- Arrange Icons is a typical Windows program option.
- The drop-down options include a list of all open windows with a check mark in front of the active window. Click on any window listed to make that window active.

10. Help

Click on the Help menu item to reveal these drop-down options.

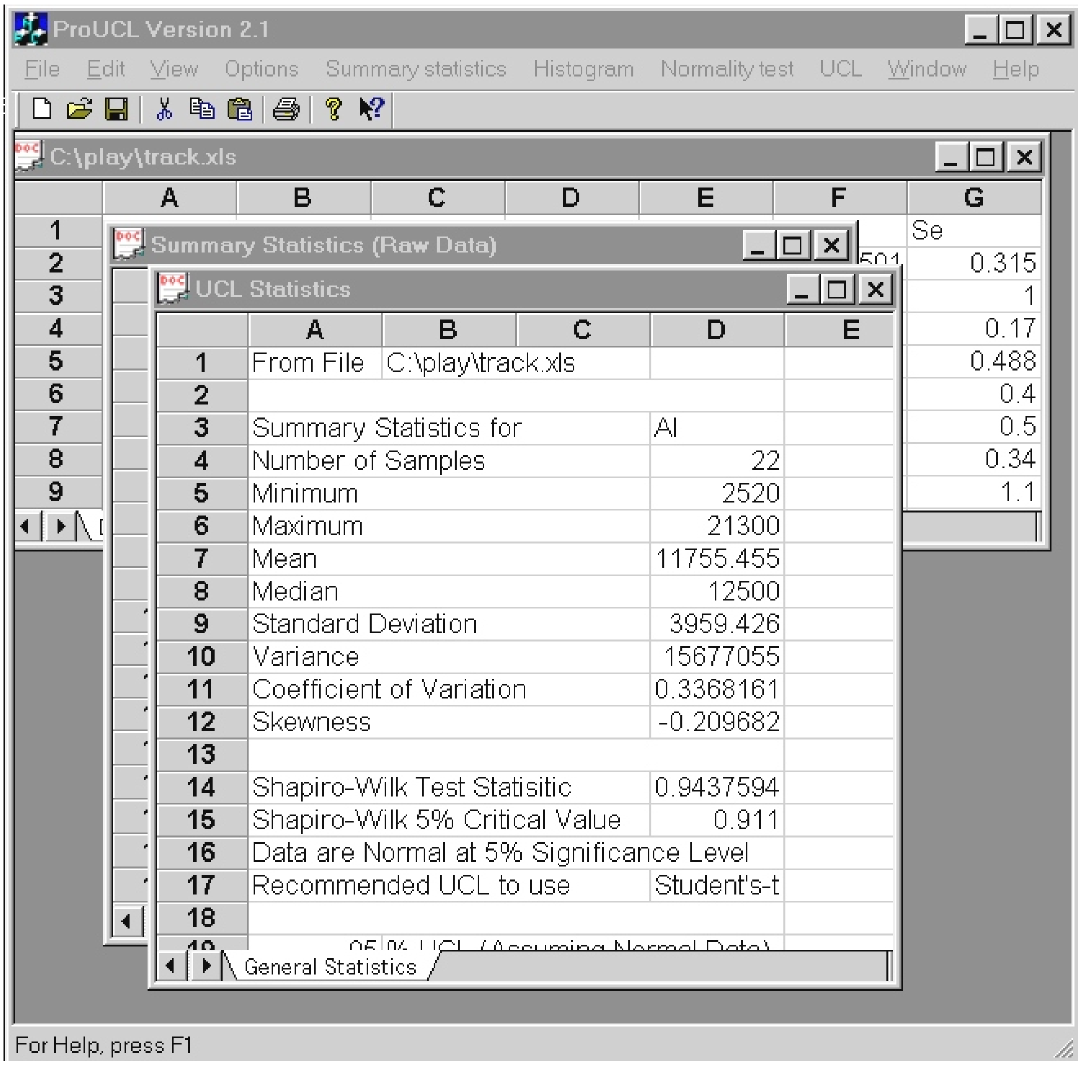


The following Help drop-down menu options are available:

- Help Topics option: at present no online help is available. This may be available in the next version of ProUCL
- About ProUCL: displays the program version number.

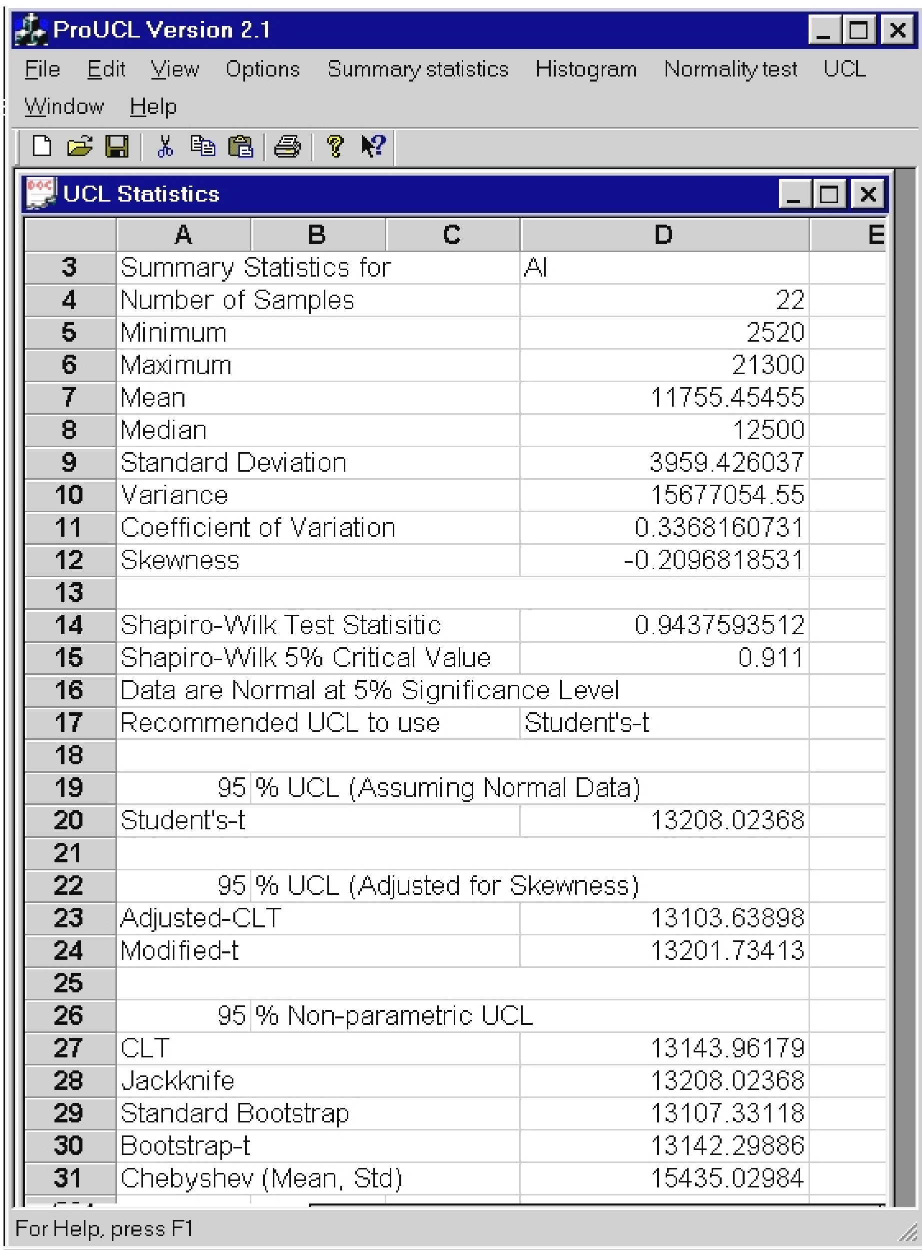
Run Time Notes

- If you have multiple windows open as shown below, you no longer need to make sure to highlight the data window before performing any computations.



- You can now do Summary Statistics, Normality Test or UCL with the screen like the above.

- Cell size can be changed. The user can change the size of a cell by moving the mouse to the top column (the grey shaded column with a letter), then moving the mouse to the right side until the cursors changes to an arrow symbol (↔), depress the left mouse button.



	A	B	C	D	E
3	Summary Statistics for			Al	
4	Number of Samples			22	
5	Minimum			2520	
6	Maximum			21300	
7	Mean			11755.45455	
8	Median			12500	
9	Standard Deviation			3959.426037	
10	Variance			15677054.55	
11	Coefficient of Variation			0.3368160731	
12	Skewness			-0.2096818531	
13					
14	Shapiro-Wilk Test Statisitic			0.9437593512	
15	Shapiro-Wilk 5% Critical Value			0.911	
16	Data are Normal at 5% Significance Level				
17	Recommended UCL to use			Student's-t	
18					
19	95	% UCL (Assuming Normal Data)			
20	Student's-t			13208.02368	
21					
22	95	% UCL (Adjusted for Skewness)			
23	Adjusted-CLT			13103.63898	
24	Modified-t			13201.73413	
25					
26	95	% Non-parametric UCL			
27	CLT			13143.96179	
28	Jackknife			13208.02368	
29	Standard Bootstrap			13107.33118	
30	Bootstrap-t			13142.29886	
31	Chebyshev (Mean, Std)			15435.02984	

- This can be used to reveal additional precision or hidden text.

Rules to remember when editing or creating a new data file.

C:\TEST\T8\UCL\ProUCL\Inorganics.xls

	A	B	C	D	E	F	G	H	I
1	Aluminum	Antimony	Arsenic	Barium	Beryllium	Cadmium	Calcium	Chromium	
2	10400	0.355	11.9	277	0.62	2.7	4160	42.6	
3	11600	0.97	6.8	86.1	0.7	0.44	2460	18	
4	11600	0.34	6.7	156	0.52	0.115	2730	22.6	
5	13600	0.465	4.7	74.8	0.65	0.045	1420	17.7	
6	12500	1.1	5.4	200	0.65	0.59	2640	17	
7	13000	1	5.6	262	0.71	0.12	2510	15.7	
8	12800		3.8	409	0.58	0.045	1040	18	
9	10800				0.55	0.045	1350		
10	9210				0.42	0.2	4930		
11									

Data

- Text may appear in the first row only. This row has column headers (variable names) for your data.

- All alphanumeric text (including blanks, strings) appearing elsewhere (other than first row) will be treated as zero data.

- Missing data (alphanumeric text, blanks) can be set to a large value=1E31. All entries with this value will be ignored from the analysis.

- The last data entry for each column must be non-zero. The program determines the number of observations by working backwards up the data until a non-zero value is encountered. Data in each column must end with a non-zero entry as shown above otherwise that zero value will be ignored. Note that, all intermediate zero entries are treated as valid data.

- It is recommended to use the default settings of the Data location screen when working with your data.

Recommendations to Compute a 95% UCL of the Population

Mean (The Exposure Point Concentration (EPC) Term)

This section describes the recommendations on the computation of a 95% *UCL* of the unknown population arithmetic mean, μ_1 of a contaminant data distribution. These recommendations are based upon the findings of Singh, Singh, and Engelhardt (1997, 1999) and Singh et al. (2000). Recommendations have been summarized for: 1) normally distributed data sets, 2) lognormally distributed data sets, and 3) data sets which are neither normal nor lognormal (non-parametric data).

Normally Distributed Data sets

- For normally distributed data sets, a UCL based upon the Student's t-statistic provides the optimal UCL of the population mean. Therefore, for normal data sets, one should use a 95% UCL based upon Student's t-statistic.

- The *95%* UCL of the mean based upon Student's t can also be used when the *sd,* $\hat{\sigma}$ (an estimate of σ) of the log-transformed data is less than 0.5, or when the data set approximately follows a normal distribution.

Lognormally Distributed Data sets

For lognormal distributions, since skewness is a function of $\hat{\sigma}$, recommendations for the UCL computation methods (Table 1) are summarized for various values of $\hat{\sigma}$ and the sample size, n. Note that the following table is applicable to the computation of a 95% UCL of the population arithmetic mean (AM) based upon lognormally distributed data sets.

Note: $\hat{\sigma}$ represents the sd of log-transformed data.

Table 1. Summary Table for the Computation of a 95% UCL of the Unknown Mean, μ_1, of a Lognormal Population

$\hat{\sigma}$	*Sample Size, n*	*Recommendation*
$\hat{\sigma} < 0.5$	For all n (≥ 5)	Student's t or H-UCL
$0.5 \leq \hat{\sigma} < 1.0$	For all n	H-UCL
$1.0 \leq \hat{\sigma} < 1.5$	$n < 25$	95% Chebyshev (MVUE) UCL
	$n \geq 25$	H-UCL
$1.5 \leq \hat{\sigma} < 2.0$	$n<20$	99% Chebyshev (MVUE) UCL
	$20 \leq n<50$	95% Chebyshev (MVUE) UCL
	$n \geq 50$	H-UCL
$2.0 \leq \hat{\sigma} < 2.5$	$n<25$	99% Chebyshev (MVUE) UCL
	$25 \leq n < 70$	95% Chebyshev (MVUE) UCL
	$n \geq 70$	H-UCL
$2.5 \leq \hat{\sigma} < 3.0$	$n < 30$	Larger of (99% Chebyshev (MVUE) UCL, 99% Chebyshev(Mean, Std))
	$30 \leq n < 70$	Larger of (95% Chebyshev (MVUE) UCL, 95% Chebyshev(Mean, Std))
	$n \geq 70$	H-UCL
$3.0 \leq \hat{\sigma}$	n small	Needs further investigation
	$n>100$	H-UCL

<u>Data Sets Without a Discernable Distribution</u> <u>(Non-parametric)</u>

- For mild to moderately skewed data sets (e.g., $\hat{\sigma}$ in the interval (0.5, 1)), one may use a 95% Chebyshev (Mean, Std) *UCL* for the population mean, μ_1. Note $\hat{\sigma}$ is sd of log-transformed data.

- For moderate to highly skewed data sets (e.g., in $\hat{\sigma}$ the interval (1.0, 2.0)), one may use a 97.5% Chebyshev (Mean, Std) *UCL* for the population mean, μ_1.

- For highly skewed to extremely highly skewed data sets with $\hat{\sigma}$ in the interval (2.0, 3.0), one may use a 99% Chebyshev (Mean, Std) to compute a *UCL* of the population mean, μ_1.

- Extremely skewed data sets with $\hat{\sigma}$ exceeding 3.0 are not well-behaved and need further investigation. For such data sets, even a 99% Chebyshev (Mean, Std) *UCL* may fail to provide the specified coverage to the population mean. This is especially true when the sample size is small.

- It is observed that the *UCL* based upon the bootstrap-t procedure is more conservative than the *UCLs* obtained using the Student's-t modified-t, adjusted-*CLT*, and standard bootstrap methods. This procedure was not included in the Monte Carlo simulation study conducted by Singh et al. (2000). It is likely that the *UCL* based upon the bootstrap t- procedure may provide better coverage of the population mean. This procedure needs further investigation.

- It is also desirable to use other distributions, such as the Gamma and Weibull distributions, to model highly skewed data sets.

It should be pointed out that, depending upon his or her application, the user may decide to use (or not use) any of the 10 available procedures incorporated in the program, ProUCL.

APPENDIX A

TECHNICAL BACKGROUND

METHODS FOR COMPUTING THE EPC TERM ((1-α) 100%UCL) AS INCORPORATED IN THE PROGRAM ProUCL

METHODS FOR COMPUTING THE EPC TERM ((1-α) 100%UCL) AS INCORPORATED IN THE PROGRAM ProUCL

1.0 Introduction

In environmental applications of the U.S. EPA, exposure assessment and cleanup decisions are often made based upon the mean concentrations of the contaminants of potential concern. A 95% upper confidence limit (*UCL*) of the unknown population arithmetic mean (*AM*), μ_1, is often used to: estimate the exposure point concentration (EPC) term (EPA, 1992), determine the attainment of cleanup standards (EPA, 1989 and 1991), estimate background level contaminant concentrations, or compare the soil concentrations with site specific soil screening levels (EPA, 1996). It is, therefore, important to compute a reliable, conservative, and stable 95% *UCL* of the population mean using the available data.

Computation of a *(1-α) 100% UCL* of the population mean depends upon the data distribution. Typically, environmental data are positively skewed, and a default lognormal distribution (EPA, 1992) is often used to model such distributions. The H-statistic based Land's (Land 1971, 1975) *H-UCL* of the mean is used in these applications. Hardin and Gilbert (1993), and Singh, Singh, and Engelhardt (1997,1999), Singh et al. (2000), pointed out some problems associated with the use of the lognormal distribution and the *H-UCL* of the population *AM*. In practice, for skewed lognormal data sets with high standard deviation (*sd*), σ of the natural log-transformed data (e.g., σ exceeding 1.5), the *H-UCL* can become unacceptably large, exceeding the 95% and 99% data quantiles, and even the maximum observed concentration, by orders of magnitude (Singh, Singh, and Engelhardt, 1997). This is especially true for samples of small sizes with high values of σ (or its estimate, s_y). In those cases, the maximum observed concentration is used as an estimate of the *EPC* term (EPA, 1992) in exposure assessment applications.

The program, ProUCL, has been developed to test normality or lognormality of the data distribution, and to compute a conservative and stable *UCL* of the population mean. Singh, Singh, and Engelhardt (1997,1999, 2000) studied several parametric and non-parametric *UCL* computation procedures which have been included in the program, ProUCL. All mathematical algorithms and formulae used by ProUCL to compute the various statistics are summarized in this Technical Background Appendix, A. ProUCL computes the various summary statistics for raw, as well as log-transformed data. In this User's Guide and in ProUCL, log-transform (*log*) stands for the natural logarithm (*ln*) to the base e. ProUCL also computes the maximum likelihood estimates (*MLEs)* and the minimum variance unbiased estimates (*MVUEs*) of various unknown population parameters. This, of course, depends upon the underlying data distribution. Based upon the data distribution, ProUCL computes the *(1-α) 100% UCLs* of the population mean using parametric and non-parametric procedures. It is observed that the Chebyshev inequality based *UCLs* provide conservative alternatives to compute a 95% *UCL* of the mean from moderately to highly skewed lognormal data sets, and other skewed non-lognormal data sets.

At present, ProUCL does not handle non-detects and missing data. The program can be modified (e.g., in the next version of ProUCL) to incorporate procedures which can be used to compute estimates of the population mean and standard deviation, and a *UCL* of the mean for left-censored data sets with non-detects.

2.0 Procedures to Test Normality and Lognormality of a Data set

ProUCL tests the normality or lognormality of the data set using the three different procedures described below. The program tests normality or lognormality at three different levels of significance, namely, 0.01, 0.05, and 0.1. The details of these procedures can be found in the references cited.

2.1 Quantile-Quantile (Q-Q) Plot

This is a simple graphical procedure to test for approximate normality or lognormality of a data distribution (Hoaglin, Mosteller, and Tukey (1983), Singh (1993)). A linear pattern displayed by the bulk of the data suggests approximate normality or lognormality of the data distribution. For example, a high value (e.g., 0.95 or greater) of the correlation coefficient of the linear pattern suggests approximate normality (or lognormality) of the data set under study. On this graphical display, observations well separated from the linear pattern displayed by the bulk data represent the outlying observations. The graphical Q-Q plot test should always be accompanied by other more powerful tests, such as the Shapiro-Wilk test or the Lilliefors test. The program ProUCL always performs the graphical Q-Q plot test on raw data as well as on standardized data.

2.2 Shapiro-Wilk W Test

This is a powerful test and is often used to test the normality or lognormality of the data distribution under study (Gilbert, 1987). The program ProUCL, performs this test for samples of size 50 or smaller. Based upon the selected level of significance and the computed test statistic, ProUCL also informs the user if the data are normally (or lognormally) distributed. The user should use this information to obtain an appropriate *UCL* of the mean. The program prints the relevant statistics on the Q-Q plot of the data (or the standardized data). For convenience, the normality (or lognormality) test results at 0.05 level of significance are also displayed on the *UCL* output Excel summary sheet.

2.3 Lilliefors Test

This test is particularly useful for data sets of larger size (Dudewicz and Misra, 1988). ProUCL performs this test for samples of sizes up to 1000. Based upon the selected level of significance and the computed test statistic, ProUCL also informs the user if the data are normally (or lognormally) distributed. The user should use this information to obtain an appropriate *UCL* of the mean. The program prints the relevant statistics on the Q-Q plot of data (or standardized data). For convenience, the normality (or lognormality) test results are also displayed on the *UCL* output Excel summary sheet.

3.0 Data

Let $x_1, x_2, \ldots, x_n$ be a random sample from the underlying population (e.g, remediated part of a site) with unknown mean, μ_1, and variance, σ_1^2. Let μ and σ represent the population mean and the population standard deviation *(sd)* of the log-transformed (natural log to the base e) data. Let $\bar{y}$ and s_y (= $\hat{\sigma}$) be the sample mean and sample *sd*, respectively, of the log-transformed data, $y_i = \ln(x_i)$; $i = 1, 2, \ldots, n$. Specifically, let

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \tag{1}$$

$$\hat{\sigma} = s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \ . \tag{2}$$

Similarly let $\bar{x}$ and s_x be the sample mean and *sd* of the raw data , $x_1, x_2, \ldots, x_n$, obtained by replacing ***y*** by ***x*** in equations (1) and (2), respectively. In this User's Guide, irrespective of the underlying distribution, μ_1, and σ_1^2 represent the mean and variance of the random variable X (in original units), whereas μ and σ^2 are the mean and variance of its logarithm, given by Y=ln(X).

4.0 Lognormal Distribution and Parameters of Interest

If $Y = \ln(X)$ is normally distributed with mean μ and variance σ^2, X is said to be lognormally distributed with parameters μ and σ^2 and is denoted by LN(μ, σ^2) . It should be noted that μ and σ^2 are not the mean and variance of the lognormal random variable, X, but they are the mean and variance of the log-transformed random variable Y, whereas μ_1, and σ_1^2 represent the mean and variance of X. The parameters of interest of a two-parameter lognormal distribution, LN(μ, σ^2), are given as follows:

Mean $= \mu_1 = \exp(\mu + 0.5\sigma^2)$ (3)

Median $= M = \exp(\mu)$ (4)

Variance $= \sigma_1^2 = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$ (5)

Coefficient of Variation $= CV = \sigma_1/\mu_1 = \sqrt{(\exp(\sigma^2) - 1)}$ (6)

Skewness $= (CV)^3 + 3(CV)$ (7)

4.1 *MLEs* of the Parameters of Lognormal Distribution

For lognormal distributions, note that $\bar{y}$ and s_y $(=\hat{\sigma})$ are the maximum likelihood estimators (*MLEs*) of μ and σ, respectively. The MLE of any function of the parameters μ and σ^2 is obtained by simply substituting these *MLEs* in place of the parameters (Hogg and Craig, 1978, Bain and Engelhardt, 1992). Therefore, replacing μ and σ by their *MLEs* in equations (3) through (7) will result in the *MLEs* (but biased) of the respective parameters of the lognormal distribution. The program ProUCL computes all of these *MLEs* for lognormally distributed data sets.

4.2 Relationship Between Skewness and Standard Deviation, σ

Note that for a lognormal distribution, the *CV* (given by equation (6) above) and the skewness (given by equation (7)) depend only on σ. Therefore, in this User's Guide and also in the program ProUCL, the standard deviation, σ (*sd* of log-transformed variable), or its *MLE,* s_y $(=\hat{\sigma})$ has been used as a measure of skewness of lognormal and of other positively skewed data sets. The larger is the sd, the larger are the *CV* and the skewness.

For example, for a lognormal distribution: with $\sigma = 0.5$, the skewness = 1.75; with σ =1.0, the skewness = 6.185; with $\sigma = 1.5$, the skewness = 33.468; and with σ = 2.0, the skewness = 414.36. Thus, the skewness of a lognormal distribution becomes very large as σ starts approaching and exceeding 2.0.

It is observed (Singh, Singh, Engelhardt (1997), and Singh et al. (2000)) that for smaller sample sizes (such as smaller than 30), and for values of σ approaching 2.0 (and skewness approaching 414) , the use of the H-statistic based *UCL* results in impractical and unacceptably large values. The various degrees of skewness of a data set as used in ProUCL and in this User's Guide are summarized as follows.

Skewness as a Function of σ ***(or its MLE,*** $s_y = \hat{\sigma}$)

Standard Deviation	***Skewness***
$\sigma < 0.5$	*Symmetric to mild skewness*
$0.5 \le \sigma < 1.0$	*Mild Skewness to Moderate Skewness*
$1.0 \le \sigma < 1.5$	*Moderate Skewness to High Skewness*
$1.5 \le \sigma < 2.0$	*High skewness*
$2.0 \le \sigma < 3.0$	*Extremely high skewness*
$\sigma \ge 3.0$	*Not well-behaved data sets - require further investigation*

4.3 MLEs of the Quantiles of a Lognormal Distribution

For highly skewed (e.g., σ exceeding 1.5), lognormally distributed populations, the population mean, μ_1,can exceed the higher quantiles (e.g., 80%, 90%, 95%) of the distribution. Therefore, the computation of these quantiles is also of interest. This is especially true when one wants to use the *MLEs* of the higher order quantiles (e.g., 95%, 97.5% etc.) as an estimate of the EPC term. The formulae to compute these quantiles are briefly described here.

The *p*th quantile (or 100*p*th percentile), x_p, of the distribution of a random variable, *X*, is defined by the probability statement, $P(X \leq x_p) = p$. If z_p is the *p*th quantile of the standard normal random variable, *Z*, with $P(Z \leq z_p) = p$, then the *p*th quantile of a lognormal distribution is given by $x_p = \exp(\mu + z_p\sigma)$. The *MLE* of the *p*th quantile is given by

$$\hat{x}_p = \exp(\hat{\mu} + z_p\hat{\sigma}). \tag{8}$$

For example, on the average, 95% of the observations from a lognormal LN(μ, σ^2) distribution would lie below $\exp(\mu + 1.65\sigma)$. The 0.5th quantile of the standard normal distribution is $z_{0.5} = 0$, and the 0.5th quantile (or median) of a lognormal distribution is M = $\exp(\mu)$, which is obviously smaller than the mean, μ_1, as given by equation (3). Also note that the mean, μ_1, is greater than x_p if and only if $\sigma > 2z_p$. For example, when $p = 0.80$, $z_p = 0.845$, μ_1 exceeds

$x_{0.80}$, the 80th percentile if and only if $\sigma > 1.69$, and, similarly, the mean, μ_1, will exceed the 95th percentile if and only if $\sigma > 3.29$.

The program ProUCL computes the *MLEs* of the 50% (median), 90%, 95%, and 99% quantiles of a lognormally distributed data set.

4.4 MVUEs of Parameters of a Lognormal Distribution

Even though the sample *AM*, $\bar{x}$, is an unbiased estimator of the population *AM*, μ_1, it does not have the minimum variance (*MV*). The *MV unbiased estimates (MVUEs)* of μ_1 and σ_1^2 of a lognormal distribution are given as follows,

$$\hat{\mu}_1 = \exp(\bar{y}) g_n(s_y^2/2), \tag{9}$$

$$\hat{\sigma}_1^2 = \exp(2\bar{y})[g_n(2s_y^2) - g_n((n-2)s_y^2/(n-1))] \ , \tag{10}$$

where the series expansion of the function $g_n(u)$ is given in Bradu and Mundlak (1970), and Aitchison and Brown (1976). Tabulations of this function are also provided by Gilbert (1987). Bradu and Mundlak (1970) give the *MVUE* of the variance of the estimate $\hat{\mu}_1$,

$$\hat{\sigma}^2(\hat{\mu}_1) = \exp(2\bar{y})[(g_n(s_y^2/2))^2 - g_n((n-2)s_y^2/(n-1))] \ . \tag{11}$$

The square root of the variance given by equation (11) is called the standard error (*SE*) of the estimate, $\hat{\mu}_1$, given by equation (9). Similarly, a *MVUE* of the median of a lognormal distribution is given by

$$\hat{M} = \exp(\bar{y})\, g_n(-s_y^2/(2(n-1))). \qquad (12)$$

For lognormal data, ProUCL computes these *MVUEs* given by equations (9) through (12).

5.0 Methods for Computing a *UCL* of the Unknown Population Mean

The program, ProUCL, computes a *(1-α) 100 % UCL* of the population mean using the following ten procedures.

1. Student's t-statistic - assumes normality or approximate normality.

2. Modified t-statistic - for skewed distributions.

3. Central Limit Theorem (*CLT*) - a non-parametric procedure.

4. Adjusted Central Limit Theorem (Adjusted-*CLT*) - for skewed distributions.

5. Land's H-Statistic - assumes lognormality.

6. Chebyshev Theorem using the sample arithmetic mean and sd (denoted by Chebyshev (Mean, Std)) - a non-parametric procedure.

7. Chebyshev Theorem using the *MVUE* of the parameters of a lognormal distribution (denoted by Chebyshev (*MVUE*)) - assumes lognormality.

8. Jackknife procedure - a non-parametric procedure.

9. Standard Bootstrap procedure - a non-parametric procedure.

10. Bootstrap t procedure - a non-parametric procedure.

The program computes a *(1-α) 100 % UCL* (except for the *H-UCL*) of the mean for any confidence coefficient *(1-α)* value lying in the interval (0.5, 1.0). For the computation of the *H-UCL,* only two confidence levels, namely, 0.90 and 0.95, are supported by ProUCL. Based upon the sample size, n, skewness, and the data distribution, the program also makes recommendations on how to obtain an appropriate *95% UCL* of the unknown population mean. These recommendations are summarized in the Recommendations and Summary Section 6.0 of this appendix. The various algorithms and procedures used to compute a *(1-α) 100% UCL* of the population mean as incorporated in ProUCL are described as follows.

5.1 *(1-α) 100% UCL* of the Mean Based Upon Student's t-Statistic

The widely used well-known *Student's t-* statistic is given by,

$$t = \frac{\bar{x} - \mu_1}{s_x/\sqrt{n}}, \tag{13}$$

where $\bar{x}$ and s_x are, respectively, the sample mean and sample standard deviation obtained using raw data. If the data are a random sample from a normal population with mean, μ_1, and standard deviation, σ_1, then the distribution of this statistic is the familiar Student's t distribution with $n-1$

degrees of freedom. Let $t_{\alpha,\ n-1}$ be the upper α quantile of the Student's t distribution with $n-1$ degrees of freedom.

A *(1−α)100 % UCL* of the population mean, μ_1, is given by,

$$UCL = \bar{x} + t_{\alpha, n-1} s_x / \sqrt{n}. \tag{14}$$

For a normally (when the skewness is about 0) distributed population, equation (14) provides the best way of computing a *UCL* of the mean. It should be pointed out that even for mildly to moderately skewed data sets (e.g., when σ starts approaching and exceeding 0.5), the *UCL* given by (14) may not provide the desired coverage to the population mean. This is especially true when the sample size is smaller than 20-25 (Singh et al. 2000). The situation gets worse for higher values of the *sd*, σ, or its estimate, s_y.

5.2 *(1-α) 100% UCL* of the Mean Based Upon Modified-t Statistic for Asymmetrical Populations

Chen (1995); Johnson (1978); Kleijnen, Kloppenburg, and Meeuwsen (1986); and Sutton (1993) suggested the use of a modified t-statistic for testing the mean of a positively skewed distribution (including the lognormal distribution). The *(1 −α)100 % UCL* of the mean thus obtained is given by

$$UCL = \bar{x} + \hat{\mu}_3/(6 s_x^2 n) + t_{\alpha, n-1} s_x / \sqrt{n} \ , \tag{15}$$

where $\hat{\mu}_3$, an unbiased moment estimate (Kleijnen, Kloppenburg, and Meeuwsen, 1986) of the third central moment, is given as follows,

$$\hat{\mu}_3 = n\sum_{i=1}^{n}(x_i - \bar{x})^3/[(n-1)(n-2)]. \tag{16}$$

It should be pointed out that this modification for a skewed distribution does not perform well even for mildly to moderately skewed data sets (e.g., when σ starts approaching and exceeding 0.75). Specifically, it is observed that the *UCL* given by equation (15) may not provide the desired coverage of the population mean, μ_1, when σ starts approaching and exceeding 0.75 (Singh, et al., 2000). This is especially true when the sample size is smaller than 20-25. This small sample size requirement increases as σ increases. For example, when σ starts approaching and exceeding 1.5, the *UCL* given by equation (15) does not provide the specified coverage (e.g., 95%) even for samples as large as 100.

5.3 *(1-α) 100% UCL* of the Mean Based Upon The Central Limit Theorem

The Central Limit Theorem *(CLT)* states that the asymptotic distribution, as *n* approaches infinity, of the sample mean, $\bar{x}_n$, is normally distributed with mean, μ_1, and variance, σ_1^2/n. More precisely, the sequence of random variables given by

$$z_n = \frac{\bar{x}_n - \mu_1}{\sigma_1/\sqrt{n}} \tag{17}$$

has a standard normal limiting distribution. In practice, this means that for large sample sizes, n, the sample mean, $\bar{x}$, has an approximate normal distribution irrespective of the underlying distribution function. Since the *CLT* method requires no distributional assumptions, this is a non-parametric method.

As noted by Hogg and Craig (1978), if σ_1 is replaced by the sample standard deviation, s_x, the normal approximation for large n is still valid. This leads to the following approximate large sample non-parametric *(1-α) 100% UCL* of the mean,

$$UCL = \bar{x} + z_{\alpha} s_x / \sqrt{n}. \tag{18}$$

An often cited rule of thumb for a sample size with the *CLT* method is $n \geq 30$. However, this may not be adequate if the population is skewed, specifically when, σ starts exceeding 0.5 (Singh, Singh, Engelhardt, Nocerino (2000)). A refinement of the *CLT* approach, which makes an adjustment for skewness as discussed by Chen (1995), is given as follows.

5.4 *(1-α) 100% UCL* of the Mean Based Upon The Adjusted Central Limit Theorem (Adjusted -*CLT*)

The "*adjusted-CLT" UCL* is obtained if the standard normal quantile, z_{α} in the upper limit of equation (18) is replaced by (Chen, 1995)

$$z_{\alpha, adj} = z_{\alpha} + \frac{\hat{\kappa}_3}{6\sqrt{n}}\left(1 + 2z_{\alpha}^2\right). \tag{19}$$

Thus, the adjusted (1 −α)100 % *UCL* for the mean, μ_1 , of skewed distributions is given by

$$UCL = \bar{x} + [z_\alpha + \hat{k}_3(1 + 2z_\alpha^2)/(6\sqrt{n})]\, s_x/\sqrt{n}. \qquad (20)$$

$\hat{k}_3$, the coefficient of skewness (raw data) is given by

$$\text{Skewness (raw data) } \hat{k}_3 = \hat{\mu}_3/s_x^3, \qquad (21)$$

where $\hat{\mu}_3$, an unbiased estimate of the third moment, is given by equation (16).

As with the modified-t *UCL*, it is observed that this adjusted-*CLT UCL* may not provide adequate coverage to the population mean when the population is skewed, specifically when σ starts approaching and exceeding 0.75 (Singh, Singh, Engelhardt, Nocerino (2000)). This is especially true when the sample size is smaller than 20-25. This small sample size requirement increases as σ increases. For example, when σ starts approaching and exceeding 1.5, the *UCL* given by equation (20) does not provide the specified coverage (e.g., 95%), even for samples as large as 100.

Thus, the *UCLs* based upon these skewness adjusted methods, such as the Johnson's modified t and Chen's adjusted *CLT*, do not provide the specified coverage to the population mean for mildly to moderately skewed (e.g., σ in (0.5, 1.0)) data sets, even for samples as large as 100. The coverage of the population mean by these *UCLs* gets worse for highly skewed data sets.

5. 5 *(1-α) 100% UCL* of the Mean Based Upon the H-Statistic *(H-UCL)*

The one-sided *(1−α)100% UCL* for the mean, μ_1, of a lognormal distribution as derived by Land (1971, 1975) is given as follows:

$$UCL = \exp\left(\bar{y} + 0.5 s_y^2 + s_y H_{1-\alpha}/\sqrt{(n-1)}\right) \tag{22}$$

Tables of H-statistic values can be found in Land (1975) and also in Gilbert (1987). Theoretically, when the population is lognormal, Land (1971) showed that the *UCL* given by equation (22) possesses optimal properties and is the uniformly most accurate unbiased confidence limit. However, it is noticed that in practice, the H-statistic based results can be quite disappointing and misleading especially when the data set consists of outliers, or is a mixture from two or more distributions (Singh, Singh, and Engelhardt, 1997, 1999). Even a minor increase in the sd, s_y, drastically inflates the *MVUE* of μ_1 and the associated *H-UCL*. The presence of low as well as high data values increases the *sd,* s_y, which in turn inflates the *H-UCL*. Furthermore, it is observed (Singh, Singh, Engelhardt, and Nocerino, 2000) that for smaller sample sizes (smaller than 15-25), and for values of σ approaching 1.0 and higher (for moderately skewed to highly skewed data sets), the use of H-statistic based *UCL* results in impractical and unacceptably large *UCL* values.

5.6 *(1-α) 100% UCL* of the Mean Based Upon The Chebyshev Theorem (Using the Sample Mean and Sample Sd)

The two-sided Chebyshev theorem (Hogg and Craig, 1978) states that given a random variable, X, with finite mean and standard deviation, μ_1 and σ_1, we have

$$P\left(-k\sigma_1 \leq X - \mu_1 \leq k\sigma_1\right) \geq 1 - 1/k^2. \tag{23}$$

This result can be applied on the sample mean, $\bar{x}$, to obtain a conservative *UCL* for the population mean, μ_1. For example, if the right side of equation (23) is equated to 0.95, then $k = 4.47$, and $UCL = \bar{x} + 4.47\sigma_1/\sqrt{n}$ is a conservative 95% upper confidence limit for the population mean. Of course, this would require the user to know the value of σ_1. The obvious modification would be to replace σ_1 with the sample standard deviation, s_x, but since this is estimated from data, the result is no longer guaranteed to be conservative. In general the following equation can be used to obtain a *(1-α) 100% UCL* of the population mean,

$$UCL = \bar{x} + \sqrt{(1/\alpha)}s_x/\sqrt{n}. \tag{24}$$

A slight refinement of equation (24) is given (suggested by S. Ferson) as follows,

$$UCL = \bar{x} + \sqrt{((1/\alpha) - 1)}s_x/\sqrt{n}. \tag{25}$$

The program, ProUCL, computes the Chebyshev *(1-α) 100% UCL* of the population mean using equation (25). This *UCL* is denoted by *Chebyshev (Mean, Std)* on the output of the program, ProUCL. Since this Chebyshev method requires no distributional assumptions about the data set under study, this is a non-parametric procedure. This *UCL* may be used as an estimate of the upper confidence limit of the population mean when data are neither normal nor lognormal, especially when sd, σ (or its estimate, s_y) starts approaching and exceeding 1.5. Recommendations on its use to a compute an estimate of the EPC term are summarized in Section 6.

5.7 *(1-α) 100% UCL* of the Mean of a Lognormal Population Based Upon the Chebyshev Theorem (Using the *MVUE* of Mean and its Standard Error)

The program ProUCL uses equation (23) on the *MVUEs* of lognormal mean and *sd* to compute a *UCL* (denoted by *(1-α)100 % Chebyshev* (*MVUE)*) of the population mean of a lognormal population. In general, if μ_1 is an unknown mean, $\hat{\mu}_1$ is an estimate, and $\hat{\sigma}(\hat{\mu}_1)$ is an estimate of the standard error of $\hat{\mu}_1$, then the following equation,

$$UCL = \hat{\mu}_1 + ((1/\alpha) - 1)^{1/2} \ \hat{\sigma}(\hat{\mu}_1) \tag{26}$$

will give a *(1-α) 100 % UCL* for μ_1, which should tend to be conservative, but this is not assured. For example, for a lognormally distributed data set, a 95% (with α =0.05) Chebyshev *(MVUE) UCL* of the mean can be obtained using the following equation,

$$UCL = \hat{\mu}_1 + (4.359) \ \hat{\sigma}(\hat{\mu}_1), \tag{27}$$

where, $\hat{\mu}_1$ and $\hat{\sigma}(\hat{\mu}_1)$ are given by equations (9) and (11), respectively. Thus for lognormally distributed data sets, the program, ProUCL, uses equation (26) to compute a *(1-α) 100% Chebyshev (MVUE) UCL* of mean. It should be noted that for lognormally distributed data sets, some recommendations to compute a 95% *UCL* of population mean are summarized in Table A1 of the Recommendations and Summary Section 6.0.

(1-α) 100% UCL of the Mean Using the Jackknife and Bootstrap Procedures

Bootstrap and jackknife procedures as discussed by Efron (1981, 1982) are nonparametric statistical techniques which can be used to reduce the bias of point estimates and construct approximate confidence intervals for parameters, such as the population mean. These two procedures require no assumptions regarding the statistical distribution (e.g., normal or lognormal) for the underlying population, and can be applied to a variety of situations no matter how complicated.

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a population with an unknown parameter, θ $(e.g., \theta = \mu_1)$, and let $\hat{\theta}$ be an estimate of θ, which is a function of all n observations. For example, the parameter, θ, could be the population mean, and a reasonable choice for the estimate, $\hat{\theta}$, might be the sample mean, $\bar{x}$. Another choice for $\hat{\theta}$ is the *MVUE* of a mean of a lognormal population, especially when dealing with lognormal data sets.

5.8 (1-α) 100% UCL of the Mean Based Upon the Jackknife Procedure

In the jackknife approach, n estimates of θ are computed by deleting one observation at a time (Dudewicz and Misra (1988)). Specifically, for each index, i, denote by $\hat{\theta}_{(i)}$, the estimate of θ (computed similarly as $\hat{\theta}$) when the ith observation is omitted from the original sample of size n, and let the arithmetic mean of these estimates be given by

$$\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{(i)}. \tag{28}$$

A quantity known as the ith "pseudo-value" is defined by

$$J_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}. \tag{29}$$

The jackknife estimator of θ is given by

$$J(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} J_i = n\hat{\theta} - (n-1)\tilde{\theta}. \tag{30}$$

If the original estimate $\hat{\theta}$ is biased, then under certain conditions, part of the bias is removed by the jackknife procedure, and an estimate of the standard error of the jackknife estimate, $J(\hat{\theta})$, is given by

$$\hat{\sigma}_{J(\hat{\theta})} = \sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}\left(J_i - J(\hat{\theta})\right)^2}. \tag{31}$$

Next, consider the t-type statistic given by

$$t = \frac{J(\hat{\theta}) - \theta}{\hat{\sigma}_{J(\hat{\theta})}}. \tag{32}$$

The t-type statistic given by (32) has an approximate Student's t distribution with $n-1$ degrees of freedom, which can be used to derive the following approximate *$(1-\alpha)100\%$ UCL* for θ,

$$UCL = J(\hat{\theta}) + t_{\alpha,n-1}\hat{\sigma}_{J(\hat{\theta})}. \tag{33}$$

If the sample size, n, is large, then the upper α t-quantile in equation (33) can be replaced with the corresponding upper αth standard normal quantile, z_α. Observe, also, that when $\hat{\theta}$ is the sample mean, $\bar{x}$, then the jackknife estimate is also the sample mean, $J(\bar{x}) = \bar{x}$, and the estimate of the standard error given by equation (31) simplifies to $s_x/n^{1/2}$, and the *UCL* in equation (33) reduces to the familiar *t*- statistic based *UCL* given by equation (14). The program ProUCL uses the jackknife estimate as the sample mean leading to $J(\bar{x}) = \bar{x}$, which in turn translates equation (33) to the *UCL* given by equation (14).

5.9 *(1-α) 100% UCL* of the Mean Based Upon Standard Bootstrap Procedure

In the bootstrap procedure, repeated samples of size *n* are drawn with replacement from a given set of observations. The process is repeated a large number of times (e.g., 2000), and each time an estimate, $\hat{\theta}_i$, of θ is computed. The estimates thus obtained are used to compute an estimate of the standard error of $\hat{\theta}$. A description of the bootstrap method, illustrated by application to the population mean, μ_1, and the sample mean, $\bar{x}$, is given as follows:

Step 1. Let $(x_{i1}, x_{i2}, \dots, x_{in})$ represent the i^th^ sample of size *n* with replacement from the original data set $(x_1, x_2, \dots, x_n)$. Then compute the sample mean and denote it by $\bar{x}_i$.

Step 2. Perform Step 1 independently *N* times (e.g., 1000-2000), each time calculating a new estimate. Denote those estimates by $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n$. The bootstrap estimate of the population mean is the arithmetic mean, $\bar{x}_B$, of the *N* estimates $\bar{x}_i : i = 1, 2, \dots, N$. The bootstrap estimate of the standard error of the estimate, $\bar{x}$, is given by,

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\bar{x}_i - \bar{x}_B\right)^2}. \tag{34}$$

If some parameter, θ (say, a population median), other than the mean is of concern, with an associated estimate (e.g., the sample median), then the same steps described above could be applied with the parameter and its estimate used in place of μ_1 and $\bar{x}$. Specifically, the estimate, $\hat{\theta}_i$, would be computed, instead of $\bar{x}_i$, for each of the N bootstrap samples. The general bootstrap estimate, denoted by $\bar{\theta}_B$, is the arithmetic mean of the N estimates. The difference, $\bar{\theta}_B - \hat{\theta}$, provides an estimate of the bias of the estimate, $\hat{\theta}$, and the bootstrap estimate of the standard error of $\hat{\theta}$ is

$$\hat{\sigma}_B = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i - \bar{\theta}_B\right)^2}. \tag{35}$$

The $(1-\alpha)100\%$ standard bootstrap UCL for θ is given by

$$UCL = \hat{\theta} + z_\alpha \hat{\sigma}_B. \tag{36}$$

The program ProUCL computes the standard bootstrap UCL by using the population AM and sample AM, respectively given by μ_1 and $\bar{x}$. It is observed that the UCL obtained using the standard bootstrap procedure is quite similar to the UCL obtained using the Student's t-statistic as given by equation (14), and, as such, does not adequately adjust for skewness.

5.10 *(1-α) 100% UCL* of the Mean Based Upon Bootstrap t Procedure

Another variation of the bootstrap method, called the "bootstrap *t*" by Efron (1982), is a nonparametric procedure which uses the bootstrap methodology to estimate quantiles of the pivotal quantity t- statistic given by equation (13). Rather than using the quantiles of Student's t-statistic, Hall (1988) proposed to compute estimates of the quantiles of statistic given by equation (13) directly from the data.

Specifically, in Steps 1 and 2 described above, if $\bar{x}$ is the sample mean computed from the original data, and $\bar{x}_i$ and $s_{x,i}$ are the sample mean and sample standard deviation computed from the *i*th resampling of the original data, the *N* quantities $t_i = (\sqrt{n})(\bar{x}_i - \bar{x})/s_{x,i}$ are computed and sorted, yielding ordered quantities $t_{(1)} \le t_{(2)} \le \cdots \le t_{(N)}$. The estimate of the lower αth quantile of the pivotal quantity in equation (13) is $t_{\alpha, B} = t_{(\alpha N)}$. For example, if $N = 1000$ bootstrap samples are generated, then the 50th ordered value, $t_{(50)}$, would be the bootstrap estimate of the lower 0.05th quantile of the pivotal quantity in equation (13). Then a *(1-α) 100% UCL* of population mean based upon the bootstrap-t procedure is given by

$$UCL = \bar{x} - t_{(\alpha N)} s_x / \sqrt{n}. \tag{37}$$

The program, ProUCL, computes the Bootstrap-t *UCL* based upon the quantiles obtained using the sample mean, $\bar{x}$. It is observed that the *UCL* based upon the bootstrap-t procedure is more conservative than the other *UCLs* obtained using the Student's t, modified t, adjusted *CLT*, and the standard bootstrap procedures. This is specially true for skewed data sets. This procedure seems to adjust for skewness. However, this procedure was not included in the Monte Carlo simulation study conducted by Singh, Singh, Engelhardt, and Nocerino (2000). This procedure needs further investigation to study the coverage probabilities provided by the *UCL* based upon the bootstrap-t method.

Note: For lognormally distributed data sets, one may want to use the jackknife and the standard bootstrap methods on the *MVUE* of the population mean, μ_1, given by equation (9). However, these procedures are not included in the program, ProUCL.

6.0 Recommendations and Summary

This section describes the summary and recommendations on the computation of a 95% *UCL* of the unknown population arithmetic mean, μ_1, of a contaminant data distribution. These recommendations are based upon the findings of Singh, Singh, and Engelhardt (1997, 1999), and Singh et al. (2000). Recommendations have been summarized for: 1) normally distributed data sets, 2) lognormally distributed data sets, and 3) data sets which are neither normal nor lognormal (non-parametric data).

6.1 Recommendations to Compute a 95% *UCL* of the Population Mean, μ_1

6.1.1 Normally Distributed Data sets

- For normally distributed data sets, a *UCL* based upon the Student's t-statistic as given by equation (14) provides the optimal *UCL* of the population mean. Therefore, for normally distributed data sets, one should use a 95% *UCL* based upon Student's t-statistic.

- The *95% UCL* of mean given by equation (14) based upon Student's t can also be used when the *sd* of the log-transformed data is less than 0.5, or when the data set approximately follows a normal distribution.

6.1.2 Lognormally Distributed Data sets

For skewed data sets, there is no simple solution to compute a *UCL* of the population mean, μ_1. Singh et al. (2000) noted that the *UCLs* based upon the skewness adjusted methods, such as the Johnson's modified - t and Chen's adjusted *CLT*, do not provide the specified coverage (e.g., 95%) to the population mean even for mildly to moderately skewed (e.g., σ in (0.5, 1.0)) data sets for samples as large as 100. The coverage of the population mean by these *UCLs* gets poorer (much smaller than the specified coverage) for highly skewed data sets as defined in Section 4.2.

For lognormally distributed data sets with a standard deviation (sd), σ, exceeding 1.0 (for moderately to highly skewed data), the use of Land's H-statistic results in unacceptably large *UCL* values (*H-UCL*), especially for samples of small sizes (e.g., smaller than 20-25). Note that even a small increase in the *sd,* σ *,* increases skewness considerably (equations (6) and (7)). For example, for a lognormal distribution, when $\sigma = 2.5$, skewness ~ 11825.1; and when $\sigma = 3$, skewness ~ 729555. In practice, the occurrence of such highly skewed data sets (e.g., $\sigma \geq 3$) is not very common. Nevertheless, these highly skewed data sets can arise occasionally and, therefore, require separate attention. Singh et al. (2000) observed that when the *sd*, σ, starts approaching 2.5 (that is, for lognormal data, when $CV > 22.74$ and skewness > 11825.1), even a *99% Chebyshev (MVUE) UCL* fails to provide the desired 95% coverage for the population mean, μ_1. This is especially true when the sample size is small (<20-25). For such extremely skewed data sets, the larger of the two *UCLs*: the *99% Chebyshev (MVUE) UCL* and the non-parametric *99% Chebyshev (Mean, Std) UCL*, may be used as an estimate of the EPC term. Another candidate to use as an estimate of the EPC term is the *UCL* based upon Bootstrap-t procedure. These issues need further investigation. It is also desirable to study other distributions such as a Gamma distribution to model the highly skewed environmental data sets.

It is also noted that, as the sample size increases, the *H-UCL* starts behaving in a stable manner. Therefore, depending upon the *sd,* σ *(actually its MLE* $\hat{\sigma}$*)*, for lognormally distributed data sets, one can always use the *H-UCL* for samples of larger sizes (e.g., 50-70 or larger). This large sample size requirement increases as the sd, $\hat{\sigma}$, increases, as can be seen in Table A1. The program, ProUCL, can compute an *H-UCL* for samples of sizes up to 1000. For lognormally distributed data sets of smaller sizes, some alternative procedures to compute a 95% *UCL* of the population mean are summarized in the following Table A1.

For lognormal distributions, since skewness (as defined in Section 4.2) is a function of σ, recommendations for the computation of the *UCL* of the population mean are summarized in Table A1 for various values of the MLE $\hat{\sigma}$ of σ and the sample size, n. Here $\hat{\sigma}$ is ML estimate of σ, and is given by the sd of log-transformed data. Note that the following table is applicable to the computation of a 95% *UCL* of the population *AM* based upon lognormally distributed data sets.

Table A1. Summary Table for the Computation of a 95% UCL of the Unknown Mean, μ_1 of a Lognormal Population

$\hat{\sigma}$	*Sample Size, n*	*Recommendation*
$\hat{\sigma} < 0.5$	For all n (≥ 5)	Student's t or H-UCL
$0.5 \leq \hat{\sigma} < 1.0$	For all n	H-UCL
$1.0 \leq \hat{\sigma} < 1.5$	$n < 25$	95% Chebyshev (MVUE) UCL
	$n \geq 25$	H-UCL
$1.5 \leq \hat{\sigma} < 2.0$	$n<20$	99% Chebyshev (MVUE) UCL
	$20 \leq n<50$	95% Chebyshev (MVUE) UCL
	$n \geq 50$	H-UCL
$2.0 \leq \hat{\sigma} < 2.5$	$n<25$	99% Chebyshev (MVUE) UCL
	$25 \leq n < 70$	95% Chebyshev (MVUE) UCL
	$n \geq 70$	H-UCL
$2.5 \leq \hat{\sigma} < 3.0$	$n < 30$	Larger of (99% Chebyshev (MVUE) UCL, 99% Chebyshev(Mean, Std))
	$30 \leq n <70$	Larger of (95% Chebyshev (MVUE) UCL, 95% Chebyshev(Mean, Std))
	$n \geq 70$	H-UCL
$3.0 \leq \hat{\sigma}$	n small	Needs further investigation
	$n>100$	H-UCL

6.1.3 Data sets Without a Discernable (non-parametric) Distribution

- For non-parametric mildly to moderately skewed data sets (e.g., σ or its estimate, $\hat{\sigma}$ in the interval (0.5, 1)), one may use a *95% Chebyshev (Mean, Std) UCL* for the population mean, μ_1.

- For populations which are neither normal nor lognormal, for moderately to highly skewed data sets (e.g., $\hat{\sigma}$ in the interval (1.0, 2.0)), one may use a *97.5% Chebyshev (Mean, Std) UCL* of the population mean, μ_1 , to obtain an estimate of the EPC term.

- For highly skewed to extremely highly skewed data sets with $\hat{\sigma}$ in the interval (2.0, 3.0), one may use a *99% Chebyshev (Mean, Std)* to compute a *95% UCL* of the population mean, μ_1.

- Extremely skewed data sets with σ exceeding 3.0, are badly behaved and need further investigation. It should be noted that for an extremely skewed data set, even a Chebyshev inequality based 99% *UCL* of the mean fails to provide the desired coverage (e.g., 0.95) of the population mean. Thus, a Chebyshev inequality based *UCL* may not be used to estimate the EPC term for data sets which are extremely highly skewed with σ approaching and exceeding 3.0.

- It is observed that the *UCL* based upon the non-parametric bootstrap-t procedure is more conservative (larger) than the other *UCLs* obtained using the Student's t, modified t, adjusted *CLT*, and standard bootstrap procedures. This is specially true for skewed data sets. The non-parametric bootstrap-t procedure was not included in the Monte Carlo simulation study conducted by Singh et al. (2000). It is likely that the *UCL* based upon the bootstrap-t procedure may provide better coverage to the population mean. This procedure needs further investigation.

- It is also desirable to study other distributions, such as a Gamma distribution, to model the highly skewed environmental data sets.

6.2 Summary of the Procedure to Compute a 95% *UCL* of Population Mean

- The first step in computing a *UCL* of a population arithmetic mean is to test for the data distribution, such as normality or lognormality of the data set. ProUCL has three procedures to test for normality: the graphical test based upon a Q-Q plot, the Lilliefors test, and the Shapiro-Wilk W test.

- ProUCL generates a quantile-quantile (Q-Q) plot to graphically test the normality or lognormality of the data. On this graph, a linear pattern displayed by data suggests approximate normality or lognormality. On this graph, points well-separated from the majority of data are potential outliers.

- After performing the normality test, ProUCL informs the user about the data distribution (normal or lognormal).

- For a normally distributed (or approximately normally distributed) data set, the user is advised to use Student's-t distribution based *UCL* of the mean.

- For lognormal data sets, the program, ProUCL, recommends (as summarized in Table A1, Section 6.1) a procedure to obtain a 95% *UCL* based upon the sample size and standard deviation of the log-transformed data, $\hat{\sigma}$. ProUCL can compute a H-*UCL* for samples of size up to 1000.

- Non-parametric *UCL* computation methods such as the modified-t, *CLT* method, adjusted *CLT* method, bootstrap and jackknife procedures are also included in the program, ProUCL. However, it is noted that non-parametric *UCLs* based upon these procedures do not provide adequate coverage to the population mean for moderately skewed to highly skewed data sets (Singh et al., 2000).

- For data sets which are neither normal nor lognormal, a non-parametric *UCL* of the mean based upon the Chebyshev theorem is preferred. The *Chebyshev (Mean, Std) UCL* does not depend upon distributional assumptions and can be used for moderately to highly skewed data sets which are neither normal nor lognormal.

- It should be noted that for extremely skewed data sets (e.g., with $\hat{\sigma}$ exceeding 3.0), even a Chebyshev inequality based 99% *UCL* of the mean fails to provide the desired coverage (e.g., 0.95) of the population mean. A procedure to compute the EPC term based upon the *Chebyshev*

(Mean, Std) UCL is described in the recommendation Section 6.1. It should be pointed out that depending upon his or her application, the user may decide to use (or not use) any of the 10 available procedures incorporated in the program, ProUCL.

References

Aitchison, J., and Brown, J.A.C. (1976), *The Lognormal Distribution*, Cambridge: Cambridge University Press.

Bain, L.J., and Engelhardt, M. (1992), *Introduction to Probability and Mathematical Statistics*, Boston: Duxbury Press.

Bradu, D., and Mundlak, Y. (1970), "Estimation in Lognormal Linear Models," *Journal of the American Statistical Association*, 65, 198-211.

Chen, L. (1995), "Testing the Mean of Skewed Distributions," *Journal of the American Statistical Association*, 90, 767-772.

Dudewicz, E.D. and Misra, S.N. (1988), *Modern Mathematical Statistics*. John Wiley, New York.

Efron, B. (1981), "Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Resampling Plans," *Biometrika*.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: SIAM.

EPA(1989), “Methods for Evaluating the Attainment of Cleanup Standards, Vol. 1, Soils and Solid Media,” Publication EPA 230/2-89/042.

EPA (1991), “A Guide: Methods for Evaluating the Attainment of Cleanup Standards for Soils and Solid Media,” Publication EPA/540/R95/128.

EPA (1992), “Supplemental Guidance to RAGS: Calculating the Concentration Term,” Publication EPA 9285.7-081, May 1992.

EPA (1996), “A Guide: Soil Screening Guidance: Technical Background Document,” Second Edition, Publication 9355.4-04FS.

Gilbert, R.O. (1987), *Statistical Methods for Environmental Pollution Monitoring*, New York: Van Nostrand Reinhold.

Hardin, J.W., and Gilbert, R.O. (1993), “Comparing Statistical Tests for Detecting Soil Contamination Greater Than Background,” Pacific Northwest Laboratory, Battelle, Technical Report # DE 94-005498.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis*. John Wiley, New York.

Hogg, R.V., and Craig, A.T. (1978), *Introduction to Mathematical Statistics*, New York: Macmillan Publishing Company.

Johnson, N.J. (1978), “Modified t-Tests and Confidence Intervals for Asymmetrical Populations,” *The American Statistician*, Vol. 73, pp.536-544.

Land, C. E. (1971), "Confidence Intervals for Linear Functions of the Normal Mean and Variance," *Annals of Mathematical Statistics*, 42, 1187-1205.

Kleijnen, J.P.C., Kloppenburg, G.L.J., and Meeuwsen, F.L. (1986), “Testing the Mean of an Asymmetric Population: Johnson’s Modified t Test Revisited.” *Commun. in Statist.-Simula.,* 15(3), 715-731.

Land, C. E. (1975), "Tables of Confidence Limits for Linear Functions of the Normal Mean and Variance," in *Selected Tables in Mathematical Statistics*, Vol. III, American Mathematical Society, Providence, R.I., 385-419.

Singh, A. (1993), “Omnibus Robust Procedures For Assessment of Multivariate Normality and Detection of Multivariate Outliers,” *Multivariate Environmental Statistics*. G. P. Patil and C.R. Rao, Editors, Elsevier Science Publishers.

Singh, A. K., Singh, Anita, and Engelhardt, M., “The Lognormal Distribution in Environmental Applications,” EPA/600/R-97/006, December 1997.

Singh, A. K., Singh, Anita, and Engelhardt, M., “Some Practical Aspects of Sample Size and Power Computations for Estimating the Mean of Positively Skewed Distributions in Environmental Applications,” EPA/600/S-99/006, November 1999.

Singh, A,. Singh, A.K., Engelhardt, M., and Nocerino, J.M. (2000), “ On the computation of the Upper Confidence Limit of the Mean of Contaminant Data Distributions.” Under EPA Review.

Sutton, C.D. (1993), “Computer -Intensive Methods for Tests About the Mean Of an Asymmetrical Distribution,” *Journal Of American Statistical Society,* Vol. 88, No. 423, pp 802-810.